THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY
of EDINBURGH

A Unified Transparency Account of Self-Knowledge

Lukas Schwengerer

PhD in Philosophy

The University of Edinburgh

2018

Abstract

In this thesis I propose an account of knowledge of one's own mental states. My goal is set on a unified transparency account of self-knowledge. It is unified, because the proposal will account for the generation of beliefs about mental states of all types, regardless of whether they are propositional, non-propositional, experiential or non-experiential. My account will thereby be applicable to knowledge of any mental state, from beliefs and desires to fears, hopes, and sensations such as pain. Moreover, it will be a transparency account because it holds on to Gareth Evans's (1982) observation that self-ascribing mental states is done by attending outwards instead of inwards. There is a sense in which we attend to the world when we find out whether we believe something, and my proposal aims to capture this intuition.

The core idea I am exploring is the following: generally, when one produces a first-order mental state, one also forms a corresponding, dispositional second-order belief about that state. Both attitudes share elements of their production, which ensures reliability while retaining fallibility. For instance, when you form a belief 'there is a red car' by perceiving a red car, you also generate the dispositional belief 'I believe that there is a red car,' if everything goes right. I argue that almost all features that make self-knowledge special can be explained with this basic idea. The assumption that the production of a first-order mental state and a second-order belief about the state go hand in hand has surprising explanatory power. Moreover, there are at least no obvious reasons why the assumption should be ruled out. The upshot will be a view that we should take seriously as a contender for an explanation of self-knowledge. I will not be able to conclusively show that it is the best explanation, but I argue that it is one worth thinking about.

The thesis is structured in three parts. The first part (chapters 1-3) focuses on the phenomenon of self-knowledge and the transparency idea. These chapters serve as the setup for my later proposed view. Chapter 1 and 2 discuss what exactly we want to explain when we say that we aim to explain self-knowledge. I thereby provide an overview of the conceptual landscape of self-knowledge and argue that we should understand the peculiarity of self-knowledge in terms of features of belief and belief-formation. Moreover, I commit myself to the view that the peculiarity has something to do with our cognitive access to mental states and relate that to the goal of a *unified* account of self-knowledge.

Chapter 3 discusses how we ought to understand the other qualification of my goal: a *transparency* account of self-knowledge. I provide an overview of transparency accounts in the literature and lay out the path to avoid common problems of transparency accounts.

In the second part (chapters 4 and 5) I propose the single process model of self-knowledge as a unified, transparency account of self-knowledge. I provide the core principles of the view and show how it explains the features of self-knowledge I aim to explain. Chapter 4 focuses on attitudes, both propositional and non-propositional. Chapter 5 expands the view to phenomenal states, such as being in pain.

The third part (chapters 6 and 7) connects the epistemological discussion of the single process model to research on cognition. Chapter 6 proposes a cognitive story of predictive processing that is compatible with the single process model. I thereby discuss the plausibility of the predictive processing idea and its empirical support. I provide a predictive processing story of self-knowledge that fits with the single process model of self-knowledge. In chapter 7 I discuss extended mental states. Clark & Chalmers (1998) propose that at least some mental states, such as beliefs, can be extended to external devices. Given that my aim is a unified account, I ought to say something about knowledge of these extended beliefs. I argue that they cannot be known by the same processes as non-extended mental states because beliefs about extended beliefs show different features than beliefs about our non-extended states that we formed by introspection. Hence, even if my view cannot account for them this is not a problem, because they are not formed by genuine introspection. Instead, we come to know extended mental states by a distinct process that we might call extended introspection.

Finally, chapter 8 provides a brief conclusion of the thesis for and points out some placed that require further development. The account is promising as an explanation of self-belief and self-knowledge, but whether it is correct also depends on future research outside the scope of philosophy.

Lay Summery

This thesis proposes an account of self-knowledge as knowledge of one's own mental states. It seems obvious in our ordinary life that we do know what we believe, desire, or intend. However, it is less obvious how exactly we can know that. If one asks me how I know that I want a piece of cake I cannot do much besides repeating that I want it. There is no other evidence I can point to and there is no other evidence that I need. I just know. There is something special about self-knowledge, such that I seem in the best position to know my own mental states. Other people have to observe my behaviour and then infer what mental state I am probably in. They might see me punch the table and conclude that I must be angry. I do not need to do any such observing. Again, apparently I just know without doing anything.

Spelling out in what way self-knowledge is special is not an easy task. One might think that the peculiarity is all about language. It seems wrong to question my claim that I believe that Vienna is in Austria, or that I am in pain. Is this a special feature of my claim that I have a certain mental state, or is there something special with my belief that I have a certain mental state? I argue that it is more than a peculiar feature of our language. Our beliefs about our own mental states are special and privileged. I have a peculiar way of forming my belief that I have a particular belief, or that I am in pain. A peculiar way only available to me that is usually reliable. Importantly, it is normally more reliable than observing another person and then making an inference to their mental state based on the observation. I might have punched the table because I saw a mosquito sitting on the table. If you observed me and concluded that I am angry you would have been wrong. On the other hand, it seems more difficult to be wrong about one's own mental states. Usually when I judge that I am angry I will be angry. Not always; sometimes my anger might be unconscious. But most of the time my beliefs about my mental states will be right. It appears to be easier to misinterpret the behaviour of another person than to misjudge one's own mental states.

The main part of my thesis develops a particular account of self-knowledge. I provide an explanation of how I generate beliefs about my mental states, such as the belief that I want a piece of cake. It is an account that can relate the belief that I want a piece of cake, to whatever generates my want for the cake. Plausibly, I want a piece of the cake because it

appears to be delicious. And the fact that it appears to be delicious should also relate to my knowledge of wanting the cake. Sometimes when someone asks me whether I believe that I want a piece of a cake I respond by thinking about the cake. How will it taste? What will the texture be? Does the cake have a soggy bottom? I figure out whether I want a piece of the cake. And in figuring out whether I want a piece of that cake I also answer the question of whether I believe that I want a piece of the cake. I come to know that I want a piece of cake by thinking about the cake. My proposed account explains this connection between thinking about the cake and figuring out whether I want the cake. Normally, both are produced together by the same process at roughly the same time. That is, whenever I form a mental state I also form a belief that I am in that mental state, if everything goes right. This basic idea explains how we can form beliefs about our own mental states without apparently doing anything in addition to forming the mental state. I form the desire to eat a piece of cake by looking at the delicious cake. At the same time I also form the belief that I desire to eat a piece of cake. Both the desire and the belief about the desire are produced in one sweep. I get the belief about my mental state for free, so to speak. However, this only happens in the case that everything goes right, and there can be a lot going wrong. We are not infallible. I might have racist beliefs, but not be aware of my racist beliefs because I also accept that racism is morally wrong. My acceptance of the moral wrongness might interfere with forming a belief that I have a racist belief. However, most of the time we are in good cases and we do know our own mental states. Hence, my account can capture our intuitions about self-knowledge. It captures how we can form beliefs about our own mental states that are often – but not always – true in a way that is different to the way we form beliefs about other people's mental states.

Acknowledgements

First and foremost I want to thank my parents Maria and Johannes who supported me throughout my studies, even with uncertain career prospects ahead of me.

I am grateful to my supervisors Aidan McGlynn and Jesper Kallestrup, who were happy to help me develop my ideas. I arrived at Edinburgh with a rather specific idea of an account of self-knowledge in mind without fully knowing whether the idea can be turned into a proper position. They helped me to transform the idea into a motivated philosophical proposal, regardless of how counterintuitive my idea initially might have looked like.

I received feedback for different papers and presentations that found their way into this thesis from a number of people to varying degree. I want to highlight some of them. Thank you to Giada Fratantonio for reading earlier versions of almost everything that is part of this thesis. Moreover, together with Kegan Shaw, Matt Jope, Michel Croce and Aidan McGlynn we formed a reading group that became the best source of feedback for my writing, and hopefully also for theirs. Thank you for being a part of that group and creating a friendly and productive environment.

I am further grateful to the 'Sperl' reading group and the 'Vienna Forum for Analytic Philosophy.' Members of these groups are partially responsible for my attempt at a philosophy PhD in the first place, so without them I would not be where I am today.

Finally, thank you to the Arts and Humanities Research Council for funding my research and taking over some of the financial burden.

Declaration

The present thesis has been composed by me and is entirely my own work. It has not been submitted for any other degree or professional qualification.

Lukas Schwengerer

19th June 2018

Contents

Introduction

I believe that this thesis is about self-knowledge. Do you trust this claim of mine? Probably. After all, I do not have any reason to start the introduction to a PhD thesis with a lie and there is no particular reason in sight why I should be wrong about myself. So you take my claim about my current belief at face value. In doing so, you grant me that I can tell you about my beliefs. You grant that I have a way of finding out what I believe right now. I agree with you. I can tell you what I believe, and I have a way of finding out what I believe. I can also tell you that I intend to finish writing this introduction, that I want a cup of tea, that I hope to get a job in academia, and that I fear that this hope will never become reality. Introspection, the acquisition of beliefs about my mental states, appears to be easy. Almost everyone can do it – or at least almost every adult human being can. Moreover, we appear to be pretty good at introspecting and thereby coming to know our own mental states. On the other hand, explaining why we can actually know our minds is difficult. What exactly is going on when I form a belief about my own mental states? How can I form true beliefs about my own mental states? I provide an attempt at an answer to these questions in this thesis. The answer ought to be not only helpful in understanding an everyday phenomenon of being able to know one's mental states, but can be useful for philosophy and psychology in a wide range of topics. Philosophers of mind who aim to understand mental states in general require an account of how we can get knowledge of our mental states. After all, they constantly rely on judgments about their own mental states when they discuss intuitions about topics such as consciousness, or the relation of beliefs and desires to action. Moreover, understanding introspection is not only a basic requirement for philosophy, but it is also relevant to psychology, which frequently relies on verbal reports from human beings regarding their own internal states. We ought to say something about our ability to report our internal states if we use such reports in further research.

In this thesis I propose a unified transparency account of self-knowledge as an explanation for the generation of self-knowledge. It is unified, because the proposal will account for the production of beliefs about mental states of all types, regardless of whether they are propositional, non-propositional, experiential or non-experiential. My account will thereby be applicable to knowledge of any mental state, from beliefs and desires to fears, hopes, and sensations such as pain. Moreover, it will be a transparency account because it holds on to Gareth Evans's (1982) observation that self-ascribing mental states is done by

attending outwards instead of inwards. There is a sense in which we attend to the world when we find out whether we believe something, and my proposal aims to capture this intuition.

The details will be developed throughout the thesis, but let me give you a glimpse of what to expect: The core idea I am exploring is that, generally, when one produces a first-order mental state, one also forms a corresponding, dispositional second-order. Both attitudes share elements of their production, which ensures reliability while retaining fallibility. For instance, when you form a belief 'there is a red car' by perceiving a red car, you also generate the dispositional belief 'I believe that there is a red car,' if everything goes right. I argue that almost all features that make self-knowledge special can be explained with this basic idea. The assumption that the production of a first-order mental state and a second-order belief about the state go hand in hand has surprising explanatory power. Moreover, there are at least no obvious reasons why the assumption should be ruled out. The upshot will be a view that we should take seriously as a contender for an explanation of self-knowledge. I will not be able to conclusively show that it is the best explanation, but I argue that it is one worth thinking about.

The thesis is structured in three parts. The first part (chapters 1-3) focuses on the phenomenon of self-knowledge and the transparency idea. These chapters serve as the setup for my proposed view. Chapter 1 and 2 discuss what exactly we want to explain when we say that we aim to explain self-knowledge. I thereby provide an overview of the conceptual landscape of self-knowledge and argue that we should understand the peculiarity of self-knowledge in terms of features of belief and belief-formation. Moreover, I commit myself to the view that the peculiarity has something to do with our cognitive access to mental states and relate that to the goal of a *unified* account of self-knowledge. Chapter 3 discusses how we ought to understand the other qualification of my goal: a *transparency* account of self-knowledge. I provide an overview of transparency accounts in the literature and lay out the path to avoid common problems of transparency accounts.

In the second part (chapters 4 and 5) I propose the single process model of self-knowledge as a unified, transparency account of self-knowledge. I provide the core principles of the view and show how it explains the features of self-knowledge I aim to explain. Chapter 4 focuses on attitudes, both propositional and non-propositional. Chapter 5 expands the view to phenomenal states, such as being in pain.

The third part (chapters 6 and 7) connects the epistemological discussion of the single process model to research on cognition. Chapter 6 proposes a cognitive story of predictive processing that is compatible with the single process model. I thereby discuss the plausibility of the predictive processing idea and its empirical support. I provide a predictive processing story of self-knowledge that fits with the single process model of self-knowledge. In chapter 7 I discuss extended mental states. Clark & Chalmers (1998) propose that at least some mental states, such as beliefs, can be extended to external devices. Given that my aim is a unified account, I ought to say something about knowledge of these extended beliefs. I argue that they cannot be known by the same processes as non-extended mental states because beliefs about extended beliefs show different features than beliefs about our non-extended states that we formed by introspection. Hence, even if my view cannot account for them this is not a problem, because they are not formed by genuine introspection. Instead, we come to know extended mental states by a distinct process that we might call extended introspection.

Finally, I want to emphasize the overall aim of the thesis that should serve as a guide to readers: I want to propose a unique idea about the relation of our production of first-order mental states and corresponding second-order beliefs and see how far it can take me in explaining self-knowledge. Philosophical research is for the most part an exercise of trial and error. This thesis is my current attempt.

1 The Phenomenon of Self-Knowledge

This chapter aims to provide a basic understanding of what the phenomenon of self-knowledge we are interested in actually is. I start with a common non-theoretic approach to set out the explanandum in part 1. In part 2, I show why this common folk starting point is less clear than is usually assumed. I further discuss two broad approaches to clarify what we seek to explain when we talk about self-knowledge in parts 3 and 4. These approaches are distinguished based on their treatment of language. On the 'linguistic view' (part 3) the peculiar features of self-knowledge are completely described by reference to our linguistic practice. On the 'doxastic view' (part 4) the peculiar features of self-knowledge are completely described by reference to beliefs and belief-formation. I give a basic characterization of both approaches and provide examples of how these views are developed in the literature. Furthermore, I sketch how an explanation of self-knowledge might look like on either approach.

1.1 Introduction

It is, excluding skeptic concerns, uncontroversial that you can know what you are thinking, that you are hungry, or that you have an appetite for biscuits and therefore intend to buy some. If a doctor at a hospital asks you where you feel pain, you have no problem to answer. If a waiter asks you for your order, you can easily say that you want the salad. However, while it seems unproblematic that you are able to testify your mental states, it is anything but clear *how* you can do that. At first sight you may believe this is just a general question about knowledge. You need to know what state you are in, and then you can report it. So you may suspect that the question *how* you can acquire knowledge is no different in cases of your own mental states than about the external world. However, at a closer look this seems to be utterly mistaken. Take the following two scenarios:

1) I claim that it is raining right now and you ask me: how do you know?
2) I claim that I am feeling cold. Again, you ask me: how do you know?

The first scenario seems not very interesting. It is an ordinary chat about the weather, including an initial claim and a question about the reasons for stating that it is raining. I can answer by citing the source – I saw it through the window. The second scenario is more interesting. There seems something wrong with questioning the mental state self-ascription. This intuitive infelicity can be explicated by noticing how a common answer to the question might look like: A reiteration of the initial claim. I know that I am feeling cold, because I *do feel cold*. I am the authority with regard to my subjectively felt temperature. I just know.

You may object that the parallel is drawn unfairly. The first scenario is about a state of the outside world and the second about a mental state. Those are two different things. However, the same intuitions seem common for one's own mental states and other's mental states.

1) I claim that Max feels cold. You ask me: how do you know?
2) I claim that I am feeling cold. Again, you ask me: how do you know?

Once again, only the second scenario seems infelicitous. The question how I know about Max's felt temperature is perfectly fine – perhaps I see his body shaking and his teeth chattering. I can give an explanation how I know and it is appropriate to demand this explanation. In the second scenario both of these are not the case. Any explanation comes down to simply restating the claim in a louder voice. Moreover, the question appears to be inappropriate.

Examples like these are the usual way to introduce the distinctiveness of self-knowledge: self-knowledge is interesting because it is different from knowledge of others.[1] A quick, non-exhaustive glance at recent work on self-knowledge shows that this is the general starting point:

[1] Note that I am here talking about self-knowledge as a paradigmatic form of knowing your own mental states. Sometimes one might get to know about one's own mental states the same way one learns about mental states of other people. That is, one might engage in a form of self-directed mind reading. In these cases one observes one's own behavior and then infers one's own mental state based on that behavior. A standard example for this type of case is Wright's explanation of a scene in Jane Austen's Emma:

> Emma has just been told of the love of her protégée, Harriet, for her — Emma's — bachelor brother-in-law, a decade older than Emma, a frequent guest of her father's, and hitherto a stable, somewhat avuncular part of the background to her life. She has entertained no thought of him as a possible husband. But now she realizes that she strongly desires that he marry no one but her, and she arrives at this discovery by way of surprise at the strength and color of her reaction to Harriet's declaration, and by way of a few minutes' reflection on that reaction. She is, precisely, not moved to the realization immediately; it dawns on her as something she first suspects and *then* recognizes as true. It *explains* her reaction to Harriet (Wright, 1998, pp. 16-17).

Crispin Wright:

> In the most salient type of case, we do not merely know ourselves best, but also *differently* from the way in which we know others and they know us. […] It remains that the type of case that sets our problem is that which gives rise to the phenomenon of avowal—the phenomenon of authoritative, non-inferential self-ascription. The basic philosophical problem of self-knowledge is to explain this phenomenon—to locate, characterize, and account for the advantage which selves seemingly possess in the making of such claims about themselves (Wright, 1998, p. 14).

Richard Moran:

> What remains before us, then, is a basic asymmetry between first-person and third-person relations. A person can make reliable psychological ascriptions to himself immediately, without needing to observe what he says or does. And this capacity lies in the nature of the first-person position itself; it is not a kind of access he may have to the mind of another person (Moran, 2001, p. 12).

David Finkelstein:

> Your friend Max tells you that he has, by mistake, bought tickets to two concerts that are taking place at the same time on the same evening. "Now," he complains, "I'll have to choose between Yo-Yo Ma and Bob Dylan." A couple of days later, you ask Max's wife if he's resolved his musical dilemma. She says that he intends to sell the Dylan ticket and attend the Yo-Yo Ma concert. That afternoon, you run into Max and congratulate him on his choice. He says, "Sarah must have misunderstood me. I may need to be out of town that weekend, in which case I'll try to sell both tickets, but if I can, I intend to see Dylan." Sarah knows Max *very* well. If you wanted to find out what size shirt he wears or how long he goes between haircuts, you'd do better to ask her than him. Nonetheless, it doesn't even occur to you to think that Max, rather than Sarah, might be mistaken about which ticket he intends to use (Finkelstein, 2003, p. 1).

Jordi Fernández:

> Thus, your justification for believing that I want Barcelona FC to win the Champions League must rely on reasoning and behavioral evidence. By contrast, I do not normally need to observe my own behavior and infer from it that I have that desire to be justified in believing that I have it (Fernández, 2013, pp. 4-5).

Brie Gertler:

> […] And when a waiter asks whether you'd like milk in your coffee, you can answer with confidence. In each case, you are in a position to arrive at knowledge of your mental state: […] It is equally clear that you are typically in a better position to identify your own mental states than others are. If others had equally good access to your thoughts, sensations, and desires, they wouldn't offer even a penny for your thoughts, or bother to consult you […] whether you want milk. So you seem to

> be *authoritative* about what you're thinking or feeling, in the sense that you can determine this by using a method that is different from, and superior to, the methods available to others (Gertler, 2011a, p. 3).

The list could go on for pages and all of them seem to share a common ground in this apparent distinction between self-knowledge and knowledge of other minds. Even deniers of an actual (i.e. not only apparent) divide feel compelled to start with the apparent immediacy. For instance, Peter Carruthers devotes a complete early chapter of *The Opacity of Mind* to the mental transparency assumption (Carruthers, 2011, Chapter 2). However, are all of them really on the same page?

Even if they all agree on a relevant distinction between self-knowledge and knowledge of others, they may not talk about the same thing after all. The difficulty arises due to the fact that the distinctive features of self-knowledge are made salient by, or even drawn from language. Just looking at the examples the focus on knowledge ascriptions and appropriate speech acts catches one's eyes. You believe Max's testimony rather than his wife's (Finkelstein). The waiter asks you whether you want milk in your coffee (Gertler). It is all about avowals (Wright) and knowledge ascriptions (Moran). This raises a general question: what exactly is the relation between language and self-knowledge?

1.2 Language as the Starting Point

It might seem natural and perhaps even obvious that we should start with looking at our ordinary language to find out what self-knowledge is. However, historically this was[2] not always the starting point (at least in mainstream European and North American philosophy) for an investigation of self-knowledge. A good way to illustrate this is to look at the beginnings of psychology as a discipline.[3] Early psychologists were interested in a way to access mental features and mental states scientifically. Introspective reports seemed like a promising approach to do so. However, early psychologists did not trust these reports to be reliable. Wundt (1888) believed that the untrained folk process of introspection is unreliable and utterly useless for scientific questions. Boring mentions that "[…] no observer who had performed less than 10,000 of these introspectively controlled reactions was suitable to provide data for published research from Wundt's laboratory." (Boring,

[2] And sometimes still is. For instance, Nichols and Stich (2003) start on the level of belief without motivating it.

[3] Interest in self-knowledge and introspection is already present long before early psychology, so this entry point is arbitrary and only meant as a useful tool to introduce the idea of language as a starting point.

1953, p. 172) To combat unreliable introspection detailed rulesets were invented to provide introspective reports that were supposed to be well calibrated and refined.[4] The important point here is that even though early psychologists developed a detailed engagement with introspection and introspective reports, the nature of speech acts of self-ascription played a surprisingly small role. Language was not used as a starting point to pick out what is special about self-knowledge. Instead, the debate started already with the assumption that there is a special kind of process that is introspection. And the peculiar nature of this process is constituted by the states that one can become aware of in virtue of introspection. What mattered to Wundt was merely how to make this process of introspection more reliable. All of their research built on the presumption that introspection as a process of detecting our mental states should be the focus. Fundamental questions whether this is the right way to understand self-ascriptions did not even arise. William James makes this explicit:

> Introspective observation is what we have to rely on first and foremost and always. The word introspection needs hardly be defined – it means, of course, looking into our own minds and reporting what we there discover. Every one agrees that we there discover states of consciousness. […] All people unhesitatingly believe that they feel themselves thinking, and that they distinguish the mental state as an inward activity or passion, from all the objects with which it may cognitively deal (James, 1890, p. 185).

Clearly, there is no discussion whether a process of detecting our mental states is the right way to think of self-ascriptions. After all, "every one agrees" on how self-knowledge comes about. In this paradigm there is no point to starting at linguistic phenomena, given that the underlying process is thought to be obvious: some kind of detection of one's mental states. Even in the late 1800s James does not even consider any other starting point than introspection as a detection of one's mental states.

The widespread preference for an ordinary language starting point is a relatively recent one. The focus on language enters towards the middle of the 20th century, led by Wittgenstein and ordinary language philosophy. The main motivation is to look at traditional philosophical problems through the lens of ordinary language in order to find new ways the problems can be solved, or dissolved. One of these problems is self-knowledge. Wittgenstein puts the focus on specific speech acts of self-ascriptions with unique features. He notes, for instance, that "[…] It can't be said of me at all (except perhaps as a joke) that I *know* I'm in pain. What is it supposed to mean - except perhaps

[4] For one list of these rules see English (1921).

that I *am* in pain?" (Wittgenstein, PI §246) Wittgenstein indicates here that the logic of self-ascriptions differs from the one governing assertions of propositions about the external world or other people. Unfortunately, he does not provide a general account of self-knowledge, even though he inspired more complete accounts (e.g. Wright (1998), Finkelstein (2003)).

Parallel to Wittgenstein's shift towards the linguistic practice of avowing Gilbert Ryle employs a similar move away from the picture of introspection as inner perception. He thereby introduces the technical term *avowal* (Ryle, 1984 (1949)) to characterize a part of what makes self-knowledge seem special. These avowals are speech acts of apparently authoritative, psychological self-ascription. They are speech acts which play a unique role in our linguistic practice. For instance, Ryle observes that it seems inappropriate to respond to an avowal such as "I want…" or "I feel hungry" with the question "Do you?" or "How do you know?" (Ryle, 1984 (1949), p. 164)[5] Introducing the term 'avowal' makes it easier to highlight the overall observation: reporting our own mental states appears to be different from assertions about other people's mental states. However, Ryle thinks that we should not take this to be an indication of any peculiar form of inner sense detecting what is going on inside our minds. This is part of his overall project to argue against the mistake to talk of the mind as if it were a physical thing. To talk of mental states as if they were things just like external objects is the origin of thinking about knowledge of these states as some sort of perception. We ought to stop making this mistake, according to Ryle. That is not to deny mental states and processes, but rather that "[…]the phrase 'there occur mental processes' does not mean the same sort of thing as 'there occur physical processes'[…]" (Ryle, 1984 (1949), p. 12). Ryle wants us to understand mental states based on our behavior, including our linguistic practice of avowing, and not as some object in a mental realm.

With Wittgenstein and Ryle the old paradigm of a perception like inner-sense becomes controversial. James's claim that "of course" introspection means looking into our mind cannot stand without justification anymore. Here enters the motivation for the linguistic starting point. We want to clearly define what the phenomenon 'self-knowledge' is. We want to determine what exactly the explanandum of our inquiry is, such that we provide a neutral starting point between the Rylean, the perceptual and any other picture that might

[5] For similar reasons Wright defines the term 'avowal' as "[…] denoting the kind of psychological self-ascription whose properties suggest privileged access" (Wright, 2015, p. 52).

be a promising candidate for explaining self-knowledge. Only then can we take the step towards providing an explanation for self-knowledge.

At this point we might still stick to the criteria of content and phenomenology to pick out self-knowledge. However, looking at language has clear advantages over both of them. A problem of the criterion of content is that it seems difficult to find an acceptably neutral starting ground on what exactly the content of self-knowledge is. The worry is that by making assumptions on the content of self-knowledge we will already decide what a potential explanation will look like. A friend of Ryle's approach will be cautious of too much of a similarity between knowledge of the external world and knowledge of mental states.

Phenomenology on the other hand is also a problematic metric. The issue here is that it is difficult to cash out the idea that there is something it is like to introspect, to acquire self-knowledge by introspection, or to have self-knowledge. How exactly should we understand this idea? It seems intuitively plausible that getting to know one's mental states seems different, but it is unclear how we can transform this into a notion that can do theoretical work. Furthermore, we cannot be sure that our judgments about our phenomenology match. That is, whether my judgments on the phenomenology of introspection matches that of any other person. This can be illustrated with the classic example of inverted qualia (or inverted spectrum), which can be traced back at least to Locke (1975 (1689)). Suppose a person A and a person B look at some object O. It seems conceivable that A has a blue color experience when looking at O, whereas B would have a green color experience looking at the same thing. Moreover, it seems also conceivable that this holds over all objects of a particular kind. So when A looks at something that B experiences at green, A would experience it as blue. However, because the inversion is so systematic their use of the color terms 'blue' and 'green' would not differ. Examples like this are often used to illustrate and defend the idea of qualia as properties of experiences that type them in phenomenological respects (e.g. Shoemaker (1982), Block (1990)). However, they also show that phenomenology is inherently subjective. To pick out a general phenomenon 'self-knowledge' we need criteria that can be used on a larger scale from a safe basis, which seems difficult with the phenomenology playing too big of a role in picking out the explanandum. Our phenomenology cannot guarantee a shared, common ground to start from. That is not to say that phenomenology cannot play any role in demarcating self-knowledge, but it should prompt us to be cautious of how much work it can do.

Language seems to be a better starting point than content and phenomenology. We have a good grasp on linguistic expressions. We largely agree which speech acts are felicitous and infelicitous. Moreover, we also largely agree which responses to these speech acts are appropriate. This makes speech acts of self-ascription a perfect, uncontroversial starting point. This is the reason that a wide range of philosophers who disagree on both explanandum and explanation of the phenomenon self-knowledge can share a common starting point in linguistic expressions of the self-ascription of mental states. Notice here that the starting point explicitly does not imply a particular explanandum. One can start with linguistic practice, but still hold that the explanandum is not identical to features of the linguistic practice. The way to establish this is to treat the linguistic practice as an indicator for something else, which then in turn is identified as the explanandum 'self-knowledge.'

The linguistic starting point consists of speech acts and responses to these speech acts. It involves different kinds of avowals (this list is a combination of Wright (1998), Bar-On (2004), and Snowdon (2012)):

a) **Experiential avowals: avowals ascribing a phenomenal state**

   "I am cold."

   "I feel tired."

   "This hurts."

   "It looks to me to be raining."

   "I am afraid of the dog."

b) **Non-experiential avowals: avowals ascribing a non-phenomenal state**

   b_1) Avowals ascribing a non-phenomenal state with intentional object

   "I am mad at John."

   "I love this song."

   b_2) Avowals ascribing a non-phenomenal, propositional attitude

   "I believe it is going to rain."

   "I hope that this paper is not terrible."

Some avowals seem to fit into both (a) and (b) depending on whether they have phenomenology or not. One might think that "I love this song" or "I am mad at John" could also be accompanied by phenomenology. Similarly, "I am afraid of the dog" might not involve a noticeable phenomenology of being afraid. Furthermore, if you accept cognitive

phenomenology, the idea that any thought has experiential qualities (cf. Pitt (2004), Bayne & Montague (2011)), the distinction between experiential and non-experiential becomes void, because any avowal of any state will have an experiential component and thereby is an experiential avowal. I take cognitive phenomenology as a strong claim about all mental states to be false, but defending this is outside of the scope of my discussion. I also bracket strong representational views of phenomenal states that blur the difference between propositional attitudes and phenomenal states, such as Tye (2000).

The linguistic starting point further involves how one can appropriately respond to these avowals. For instance, it seems inappropriate to respond to a phenomenal avowal with the question "How do you know?" Similarly, it seems infelicitous to claim that someone is wrong when she sincerely avows that she feels tired, provided that there are no reason to doubt her conceptual competency and attention.[6] On the other hand, challenging a non-experiential avowal can be appropriate even if the speaker is sincere, competent and attentive. Nevertheless, these challenges are rare occurrences in our everyday life.

This list is undoubtedly incomplete, but provides a general idea on how to think about language as a starting point. The next question is where we should go from this starting point. That is, what is the phenomenon 'self-knowledge' that we can point out by looking at these linguistic practices? There are at least two options available. One proposes that features of self-knowledge should be identified wholly with our linguistic practice of avowing; the other demands something beyond speech acts and identifies self-knowledge with features on the doxastic level. Wright formulates this choice:

> So the would-be theorist of self-knowledge confronts a fork. What comes first here in the order of explanation: the linguistic practice, or the thoughts of the thinkers manifested in that practice? The problem of self-knowledge will look very different depending on how we choose (2015, p. 52).

As Wright indicates it is a question relevant to any inquiry into self-knowledge. Even if we all agree that there is something special and privileged about linguistic practices related to self-knowledge, it is unclear where the features of self-knowledge that we want to explain are located. Moreover, this decision of locating self-knowledge determines how the features can be explained. If we take self-knowledge to be an instance of special and privileged belief we have to explain what makes it special and privileged in terms of

[6] Some philosophers think that even an experiential avowal can be appropriately challenged. See Schwitzgebel (2008).

properties of belief and belief-formation. If we locate it at the level of linguistic practice, as an instance of avowals as a special kind of speech act, we have to explain the specific rules governing the speech act. In this case we need to explain what makes the speech act special. Hence, Wright rightfully emphasizes that this choice of a starting point needs to be properly addressed. We are confronted with a choice between a *linguistic view*, and a *doxastic view*.[7]

We can capture the two options with these two principles:

(Linguistic View) The peculiar nature of self-knowledge should be completely described by features of linguistic practice, syntax, semantics, and pragmatics.

(Doxastic View) The peculiar nature of self-knowledge should be completely described by features of beliefs and belief formation.

(Linguistic View) expresses the language-first path of Wright's fork, whereas (Doxastic View) captures the thought-first path. The idea is that we need to fully describe what exactly ought to be explained at what point. If we want to follow Wright's talk of order of explanations, we need to determine what we ought to explain first. The linguistic view states that what we ought to explain first is a set of features on the level of linguistic practice. So we should start by describing these features of linguistic practice. The doxastic view in contrast states that what we ought to explain first is a set of features on the level of thought – belief and belief-formation, so we should start by describing these features of belief and belief-formation. Both views set up the next step of a discussion of self-knowledge by defining what kind of features ought to be explained.

One might be worried here about my definition of the linguistic view as *completely* describing the features. Couldn't one also understand the talk of order of explanation differently? One might merely propose that we should first explain linguistic features and then doxastic feature, but claim that self-knowledge has features on both levels. I take this conception of self-knowledge to be difficult to even conceive. The proposal appears to posit two different phenomena bundled together under a single term and claims that it is one phenomenon. It is unclear how one phenomenon can have linguistic and doxastic features. Furthermore, Wright does not seem to allow this option. As I am going to discuss in detail in

[7] This terminology is mine. A similar difference about the target explanandum has also been observed by Jongepier and Strijbos (2015).

chapter 2, Wright believes the advantage of the linguistic view is that it remains neutral on whether there is any feature on the doxastic level that is important for self-knowledge. He wants to allow for a deflationist account of self-knowledge, which denies any privileged belief about one's mental states. A deflationist account is only possible if the explanandum self-knowledge is set up without doxastic features. Otherwise the deflationist view would be ruled out already from the start. Hence, whoever wants to keep the deflationist view on the table has to describe the explanandum *completely* in non-doxastic terms. In doing so we have to bite the bullet on the term 'self-knowledge' becoming a misnomer: It will pick out a specific type of speech act and its features. Hence it differs from the term 'knowledge,' that picks out a combination of justified true belief plus an anti-Gettier condition, or a distinct mental state type. This is not to deny that there cannot be any privilege on the doxastic level. Whatever happens on the doxastic level can still be relevant. However, it is only relevant insofar as it relates to features on the linguistic level. The doxastic level itself is not part of the explanandum self-knowledge, according to the linguistic view.

In the next parts I provide an overview of positions subscribing to the linguistic view and positions accepting the doxastic view. In chapter 2 I will discuss whether one of these views should be preferred over the other.

1.3 Accepting the Linguistic View

For proponents of the linguistic view it is relatively easy to take the step towards defining the features of self-knowledge that require an explanation. Take the linguistic practice, look for consistent differences between avowals and other speech acts and generalize them into features of self-knowledge. What features of the linguistic practice one chooses for this purpose differs in literature. This might be surprising given that the linguistic practice is generally agreed upon. Different versions of the linguistic view are indeed similar to each other. However, they differ in fine grained distinctions about the linguistic practice. That is, they might start with the same folk judgments on the felicity of avowals, but understand their relation differently. Moreover, different versions of the view might also be distinguished by the extent to which certain responses to avowals are inappropriate. For instance, even though all versions of the linguistic views include that challenging avowals is to some extent inappropriate, some of them might propose that for specific classes of

avowals it is always inappropriate to challenge them, whereas other versions only hold that it is generally inappropriate, but can be appropriate in unusual circumstances.

All proponents of the linguistic view need to satisfy the requirement that the features do not involve anything besides linguistic practice, syntax, semantics, and pragmatics. They need to be careful that their definitions of the features of self-knowledge do not include features on the doxastic level. I am going to discuss three different ways in current literature of spelling out the explanandum 'self-knowledge' following the linguistic view. I start with Crispin Wright's illustration, then discuss Dorit Bar-On's work and finally sketch David Finkelstein's position.

1.3.1 Crispin Wright

Wright (1998; 2001; 2015) proposes *immediacy, authority, and salience* as the explanandum for self-knowledge.[8] These are features of avowals defined with reference to appropriate speech acts and responses. The exact nature of them differs depending on the type of avowal in question. Wright does not use the distinction of experiential and non-experiential avowal, but rather one of phenomenal and attitudinal avowal. Both categories are not explicitly defined, but rather explained by a selection of examples. Phenomenal avowals are provided with examples like 'I have a headache', 'My feet are sore', or 'I'm tired.' (Wright, 1998, p. 14) They differ with my experiential avowals insofar as phenomenal avowals do not involve attitudes at all. For instance, 'It looks to me to be raining' is not a phenomenal avowal.

Phenomenal avowals show three key features (Wright, 1998, pp. 14-15):

1. *Immediacy,* which captures the observation that demanding reasons or evidence from a person that voices an avowal seems inappropriate. The question 'How can you tell?' is always inappropriate as a response to a phenomenal avowal.
2. *Strong authority, which* means that whenever someone sincerely claims that she is in mental state x, and understands what this claim means, this guarantees that the claim is true. Any doubt about such a claim has to be a doubt about sincerity or understanding.

[8] Wright (1998; 2001) uses *groundlessness* instead of immediacy, and *transparency* instead of salience.

3. *Salience*, defined by the absurdity of avowing uncertainty about one's own mental states. If one is asked 'Do you have a headache?' it seems absurd to answer 'I don't know.'

These features of phenomenal avowals have been attacked by Snowdon (2012). For instance, he provides the following case against strong authority:

> Consider first a case relevant to authoritativeness. A teenager on amphetamines crashes a car and is hurled through that window and ends up under the car. The doctor tries to locate the injured areas in a systematic way, by first pressing the areas which do not seem damaged. However, the boy screams and claims that he is in pain. It seems to me that it would not be at all irrational for the doctor to hold that the boy was so confused and frightened and disordered that he thought that the pressure gave him pain when it did not. The doctor could think this without hypothesizing that the injured boy was either lying or did not understand English. In circumstances such as these there is no norm about avowals corresponding to Wright's concept of authoritativeness (Snowdon, 2012, pp. 251-252).

And the following experience as a counterexample to salience:

> For example, when having my eyes tested I am asked which of two lenses results in the greater blurring. It can be very hard to say, and I can aver that I do not know which is blurrier. Now, this is a comparative judgement which relies on memory, but it would not be true to the experience to suppose that the worry is generated by not remembering. The problem is, rather, that it is hard to judge which is blurrier. This seems to be a phenomenal judgement about which one can aver ignorance (Snowdon, 2012, p. 252).

Wright accepts these counterexamples to some degree. He concedes that something might obstruct the proper exercise of one's judgmental and conceptual faculties. Cases of distraction, panic, or pain might hinder salience or authority. However, Wright denies that his conditions are completely off-track. He merely concedes that they require more careful formulations. Hence, Wright (2015) proposes a clause that rule out such obstructions in a formulation of salience.[9] Similarly, he adds provisos to the claim of strong authority that rule out cases in which there are obstacles to any proper judgment. For instance, the teenager in the car crash might not be able to judge correctly whether he is in pain, because his cognitive apparatus is not working properly in a mix of amphetamines, injuries, and panic. But this is a situation in which any judgment is obstructed, so it does not show

[9] Though this formulation is spelled out in Snowdon's terms, so according to the doxastic view:

> Necessarily, if phenomenal condition C applies to S and S possesses the relevant concepts and is in no condition that would impair their exercise in judgment, then S will know that C applies to him/her. (Wright, 2015, p. 67)

that self-knowledge is not special in regular cases. It merely shows that self-knowledge also relies on some basic cognitive and conceptual capacities. Moreover, Wright does stand firm on his overall methodological point. His aim is not primarily to provide the best possible formulation of the features of self-knowledge, but rather to show how such features might look like if we define them on a level of linguistic practice. Having to refine the formulations without being forced to a level of belief is perfectly fine for his aim. Hence, Wright argues that "[…] for anyone inclined to take the 'linguistic turn' […] we have found no serious cause in Snowdon's discussion for loss of confidence" (Wright, 2015, p. 74).

The second category of speech acts of self-ascriptions are attitudinal avowals. These are avowals of content-bearing states, such as 'I believe that term ends on the 27th,' and 'I hope that noise stops soon' (Wright, 1998, p. 15). Attitudinal avowals show:

1. *Immediacy*, again referring to the observation that demanding reasons or evidence from a person that voices an avowal seems inappropriate. The question 'How can you tell?' is usually inappropriate as a response to an attitudinal avowal.
2. *Weak authority,* meaning that attitudinal avowals provide empirically assumptionless justification for the corresponding third-person claims. One can (even though one rarely does) doubt individual attitudinal avowals without doubting sincerity or understanding. However, one cannot doubt that a person correctly avows attitudes in general. Wright labels speech acts with weak authority as *inalienable*.
3. *Salience*, again referring to the absurdity of avowing uncertainty about one's own mental states. If one is asked 'Do you believe that p?' it seems absurd to answer 'I don't know.' However, it might not be absurd to answer that in the sense of suspending judgment. It would only be absurd if you did not know whether you believe that p, not believe that p, or are withholding judgement.

I find the inalienability that comes with weak authority puzzling. Moreover, Wright's motivation for this feature is equally puzzling to me. He provides the following argument:

> You may not suppose me sincere and comprehending, yet chronically unreliable, about what I hope, believe, fear, and intend. Wholesale suspicion about my attitudinal avowals—where it is not a doubt about sincerity or understanding—jars with conceiving of me as an intentional subject at all (Wright, 1998, p. 18).

I do not see an obvious reason why this kind of suspicion is in conflict with conceiving someone as an intentional subject and Wright provides no further elaboration. Regardless

of whether this argument is successful, weaker authority can be seen as some kind of "[…] authority with the propriety of deference to what she has to avow […]," as Wright (2015) states. Whenever someone avows that she believes p, we trust that she is right. We have a presumption that subjects tell the truth, even though there is no guarantee when they avow attitudes. And this description does fit the linguistic view. This authority is a feature of linguistic practice, not a feature of beliefs.

1.3.2 Dorit Bar-On

Bar-On's (2004) formulation of the explanandum 'self-knowledge' looks similar to Wright's conception. She also formulates it in terms of features of avowals which are indicated in the desiderata her account aims to explain (she states 8, I only mention the 4 relevant ones that capture epistemic asymmetry):

> D1. The account should explain what renders avowals protected from ordinary epistemic assessments (including requests for reasons, challenges to their truth, simple correction, etc.).
>
> D2. It should explain why avowals' security is unparalleled: why there are asymmetries in security between avowals and all other empirical ascriptions, including (truth-conditionally equivalent) third-person ascriptions and non-mental first-person ascriptions. In particular, it should explain why avowals are so strongly presumed to be true.
>
> D3. It should explain the non-negotiable character of the security—the fact that it is 'non-transferable' and 'inalienable'.
>
> D4. It should apply to both intentional and non-intentional avowals alike, and allow us to separate avowals from other ascriptions in terms of their security (Bar-On, 2004, p. 20).

Bar-On (2004; 2010) motivates these features by pointing to ordinary examples. In everyday cases no one questions my avowals, and if they do it usually seems inappropriate. Ordinary assessment of avowals seems out of place. Furthermore, we do have a presumption that the avowal is true. On the other hand, my assertions about external objects are comparatively frequently questioned and these challenges do not seem inappropriate at all. Moreover, if you learn about my mental states through my testimony you would still not end up with the same security that I have with regard to my mental states. The security is distinctly first-personal.

There is no feature parallel to Wright's salience idea in Bar-On. In her characterization it seems perfectly fine to respond with 'I don't know' to a question of whether one is in pain,

or whether one believes something. This might strike one as odd, partially because of phenomenological observations, and partially because of the emphasis on salience in discussions in the epistemology literature (e.g. debates on Williamson's (2000) anti-luminosity argument). Moreover, it might lead to problems accounting for negative avowals – avowals of what mental state one is not in (Brueckner, 2011). For instance, we seem to be able to make especially secure avowals of not being in pain, or not believing that p. It is at least not obvious how these avowals fit into Bar-On's framework.

In addition, Bar-On uses the same features for all kinds of avowals compared to Wright's differences between phenomenal and attitudinal avowals. The price for this is that no guarantee of truth is found in her features, whereas Wright proposed that sincerity guarantees truth for phenomenal avowals.

In rare passages Bar-On is dangerously close to bringing in features on the doxastic level into the explanandum. Hence, one might be worried that Bar-On does not deliver the non-epistemic approach she promises. For instance, she notices that avowals are issued with "a very high degree of confidence" (Bar-On, 2004, p. 3). The notion of having a high degree of confidence does not seem to fit well with the linguistic project. We have to be careful not to identify this with a high degree of confidence in an epistemic sense. It cannot be the confidence in a belief. That is, it cannot be a particularly high justification for a belief, or a certain subjective probability for the belief being true. Moreover, it cannot be a phenomenal state of having confidence that something is the case, because that would also leave the scope of the linguistic view. What should we do with the confidence claim in Bar-On's description? I take it that there are two options available. First, we can attribute it to a simple misstep in the writing that should be ignored. After all, the confidence claim does not show up in the desiderata at all. Second, we can interpret the confidence claim as a claim about linguistic expression. A confident speech act in this way should be treated as a speech act that puts additional responsibility on the speaker. Take an assertion as illustration. If I confidently assert something I thereby make it clear to the hearer that I am responsible for the truth of the assertion. On the other hand, if I assert something in a tentative fashion I thereby signal that even though I am asserting something, the hearer should be careful in believing solely on my say-so. If the hearer nevertheless does that, she thereby takes up responsibility for doing so. If it turns out that she now believes something

false she cannot fully blame me. Confidence of a speech act in this fashion is related to appropriate responses to the speech act.

Bar-On's discussion of self-knowledge comes with a further complication. While she does subscribe to the linguistic view for the most part, she does not fully commit to this position. Bar-On distinguishes three different questions related to self-knowledge.

> (i) What accounts for the *unparalleled security of avowals*? Why is it that avowals, understood as true or false ascriptions of contingent states to an individual, are so rarely questioned or corrected, are generally so resistant to ordinary epistemic assessments, and are so strongly presumed to be true?
>
> (ii) Do avowals serve to articulate *privileged self-knowledge*? If so, what qualifies avowals as articles of knowledge at all, and what is the source of the privileged status of this knowledge?
>
> (iii) Avowals aside, what allows us to possess privileged self-knowledge? That is, how is it that subjects like us are able to have privileged, non-evidential knowledge of their present states of mind, regardless of whether they avow being in the relevant states or not?
>
> (Bar-On, 2004, pp. 11-12)

As a starting point (iii) is discarded because it assumes privileged self-knowledge and thereby also denies negative answers to (i) and (ii). The methodologically interesting decision is to pick (i) over (ii). Bar-On believes that the interest in the phenomenon of self-knowledge arises from the asymmetry to knowledge of others, and this asymmetry is directly taken from the special nature of avowals in conversation and thought. We need to start at the avowals, because they determine our perspective of self-knowledge. Starting at the avowals with question (i) is the move that fits the linguistic view. The question (i) is formulated in terms of linguistic practice, and thereby the explanandum is the set of language based features D1-D4. However, she does not commit to answers to (ii) and (iii). She rather provides different possible answers, some negative, others positive. That is, Bar-On (2004, p. Chapter 9) provides some combination of these answers to (ii) and (iii) that fit with the linguistic view, and others that do not.[10] Some answers that do not fit with the linguistic view lead towards a hybrid view between linguistic and doxastic. For instance, in one option the notion of belief is weakened such that an avowal itself is enough to constitute a belief (Bar-On, 2004, p. 365). In this option the clear distinction between

[10] See also Bar-On (2011) where she reaffirms that she does not argue for any settled view on privileged self-knowledge.

linguistic and doxastic is lost. Because she discusses various different possible views it is difficult to evaluate Bar-On's overall picture of self-knowledge. Nevertheless, the special security of self-knowledge is fully understood in terms of the linguistic view. Moreover, it is clear that Bar-On does not want to rule out a negative answer to question (ii) in her initial set-up. Cautiously formulated we can say that there is at least one Bar-On *inspired* position that is fully in the spirit of the linguistic view. This is a view that understands the explanandum for (i) according to the linguistic view, and provides negative answers to (ii) and (iii).

1.3.3 David Finkelstein

Finkelstein (2003) says relatively little about what the phenomenon self-knowledge is, but he clearly understands it on a similar line to Wright and Bar-On. He introduces the topic with the case of Max I already cited earlier. Max avows that he intends to see Dylan rather than Yo-Yo Ma if he can only see one of them, whereas his wife claimed that he intends to go to the Yo-Yo Ma concert. Finkelstein notes two features about this case that characterize what the problem of self-knowledge amounts to in his view. First, one is the authority concerning one's mental states – in Max's case his intention. Finkelstein understands being the authority with regard to x as being the best person to ask about x. In this case Max is intuitively the best person to ask about his own mental states. Finkelstein does not make the notion of 'being the best person to ask' more precise. He rather takes this notion to be clear enough for our ordinary linguistic practice. However, he does add a small caveat stating that being the best person does not mean that the person cannot be wrong. He clearly endorses fallibility. Moreover, he carefully adds that the avowing person usually is the best person to ask, but might not be in some abnormal situations. Nonetheless, one might think that 'being the best person to ask' still means that one is the most likely to make a correct assertion about one's mental state. Finkelstein does not explicitly endorse this, even though it seems plausible that if someone is the best person to ask she will also be the one most likely to lead one to true beliefs.

The second feature Finkelstein mentions, although only in passing, is that "it does not occur to you to ask Max for evidence supporting his assertion that he intends to see Dylan" (Finkelstein, 2003, p. 1). Asking for evidential backing for avowals is not part of our linguistic practice. Finkelstein does not put it in terms of the question being inappropriate, but rather suggests that if one were to ask this question it would not be understood as a

straightforward prompt to provide evidential support. Rather it would be understood as some other speech act (e.g. a joke). Asking for evidential support is simply so far-fetched that one has pragmatic reasons to interpret the question as being some other speech act. It is not far-fetched because we are especially great at detecting our mental states, but rather because of how our speech acts of avowing function. The speech act is such that it can convey information without committing oneself to being able to provide evidential support. Hence Finkelstein thinks that what we want to explain is understood solely on a level of the linguistic practice: Avowals that are authoritative and come without any need of being potentially backed up with evidence.

1.3.4 Explaining Self-Knowledge

Accepting the linguistic view allows flexibility for possible explanations of the features of self-knowledge. Any explanation has to provide a convincing story how the features of our linguistic practice come about, but the linguistic view is not committed to an explanation that stays on the level of speech. The linguistic view does not commit one to deny privileged access on the level of belief. It is compatible with a story on the level of belief that explains why our linguistic practice of avowing exists in its current form. So one might argue for higher reliability of self-beliefs to explain why challenging avowals seems inappropriate. However, the linguistic view keeps options outside of the level of belief open. It allows non-epistemic explanations of the features of self-knowledge. Hence, it gives rise to non-epistemic accounts, such as simple expressivism, Wright's 'default view,' or Bar-On's (2004) and Finkelstein's (2003) versions of neo-expressivism.[11] The latter two build on the idea that avowals are not primarily reports of having certain mental states. Rather, avowals express mental states, whereas the relation of expressing a mental state explains the peculiar features of avowals. Both propose that the reason that avowals are especially secure is simply because you cannot make certain mistakes when avowing. Bar-On explains this with the help of the notion of *immunity to misascription*: One does not attempt to recognize mental states at all in order to avow a mental state, and therefore one cannot fail in recognizing mental states. Challenging an avowal is inappropriate,

[11] Expressivists treat avowals similar to natural expressions like grimaces or crying: they function to express a mental state of a person rather than reporting the mental state. A brute form of expressivism denies that avowals are anything besides expressions. Crucially, they are not assertions at all. (Wright, 1998, p. 34) Neo-expressivists on the other hand propose that avowals can have this expressive function, while still being truth-evaluable. Cf. Finkelstein (2003), Bar-On (2004), Bar-On & Sias (2013)

because it involves a claim that such a mistake has been made, even though this kind of mistake is impossible here. Precisely the fact that you do not engage in a specific process on the level of belief makes it inappropriate to challenge your avowal. This sort of non-epistemic explanation[12] of the features of self-knowledge is possible because the linguistic view is neutral between epistemic and non-epistemic explanations. The linguistic view leaves the realm of possible explanations wide open. I will further discuss this advantage of the linguistic view and whether it can fully capture our folk intuitions on self-knowledge in chapter 2.

1.4 Accepting the Doxastic View

To adopt the doxastic view from the starting point of language means to accept that observations of our linguistic practice can be a basis for an inference to features of beliefs about one's mental states and their formation. This way one treats avowals as nothing more than ordinary reports. They are a result of self-belief, and we can learn a lot about self-belief and self-knowledge by looking at them. Nevertheless, they themselves are not the primary part of what makes self-knowledge special and hence do not constitute the features we want to explain. The peculiarity of avowals is only derived from the peculiarity of self-beliefs, according to the doxastic view.

The idea leading to formulations of the doxastic view is that we can use observations of our linguistic practices to get a grasp on the features that our beliefs about our mental states have. We cannot look at our concepts directly, but they materialize in linguistic practices and linguistic intuitions. Therefore a good way to understand a concept is to look at how we use it in language.

1.4.1 From Language to Belief

Let us grant the connection between avowals and self-belief for now. How exactly should this connection influence our notion of self-knowledge that we want to explain? One version holds that the linguistic practice is a helpful form of evidence available to determine the features of self-knowledge and we need to respect that by putting a priority to vindicate our intuitions about our linguistic practice. Our theories should reflect these intuitions.

[12] These explanations are non-epistemic insofar as they do not "derive avowals' special security from the security of a special epistemic method, or privileged epistemic access." (Bar-On, 2004, p. 11)

An example of a methodology of this kind in epistemology can be found in the debate on norms of action (especially norms of assertion). These are, roughly, conditions under which an action (assertion) is proper. A line of argument in the debate about these norms starts with our ordinary linguistic practice of expressing evaluations of action and takes this as the prime evidence for our epistemic practice. Observing our conversational patterns shows that we use 'know' to indicate whether someone satisfies a norm of action.[13] We can use this observation as evidence that our epistemic practice also relies on the concept of knowledge. For instance, Hawthorne and Stanley (2008) emphasize the use of 'know' as evidence:

> [I]t bears emphasis that (in English at least) it is considerably more natural to appraise behavior with the verb 'know' than the phrase 'justified belief', or even 'reasonable belief'. Perhaps this is because 'know' is a phrase of colloquial English, whereas 'justified belief' is a phrase from philosophy classrooms. But this is itself a fact that should be surprising, if the fundamental concept of appraisal were justification rather than knowledge (2008, p. 573).

And similarly, Williamson (2000) takes our "[…] conversational patterns to confirm a knowledge norm of assertion" (2000, p. 252). The assumption in play here is that we can observe our epistemic practices in our linguistic practices. Conversational patterns can only confirm an epistemic norm, if these pattern themselves are sufficiently related to our epistemic practice. If we assume that our linguistic practice is sufficiently connected to our epistemic practice, we can use the linguistic practice as evidence. Moreover, we can use the linguistic practice also as a tool to develop accounts of our epistemic practice. Mikkel Gerken (2017) describes this as follows:

> Roughly, our folk epistemology consists of the tacit principles and presuppositions that underlie and guide our everyday cognitive and linguistic epistemic practices. For example, they govern our pre-theoretical epistemic assessments of one another. Since these principles and presuppositions are tacit, we must "reverse engineer" our way to an articulation of them via reflection on our everyday epistemic practices (p. 16).

Our folk epistemology guides our *linguistic* epistemic practices. Hence, we can look at our ordinary language to learn something about the tacit epistemic principles and presuppositions. However, this might not be as straightforward as one would hope. For instance, in case of the debates on norms of assertion, the mere fact that we frequently use

[13] This is often supported by pointing to studies showing how frequently we use 'knows.' According to Davies and Gardner (2010) 'knows' is among the ten most frequently used verbs in English.

'know' to express evaluations of actions does not demonstrate with certainty that we do have a knowledge norm in place. Instead, the linguistic practice is compatible with various different epistemic practices that might underlie our evaluations. Hence, Gerken (2013; 2017) is not convinced by the knowledge norm, even though he accepts the path from linguistic observations to an account of epistemic practices. Similarly, Douven (2006), Brown (2008), and Kvanvig (2009) deny the knowledge norm.[14] The lesson to learn is that even if we accept that the linguistic practice can guide us to our epistemic practice, the linguistic practice will not lead to a single, definitive account. The linguistic practice underdetermines our epistemic practice: the same linguistic practice is compatible with a multitude of epistemic practices.

Applying a similar methodology to self-knowledge is to take the intuitions regarding avowals at face value and infer the features of self-knowledge from the features of avowals. In the framework of the doxastic view the epistemic features underlie the linguistic practice, so we can – as Gerken aptly describes – reverse engineer the features of self-knowledge on a doxastic level. When avowals show some sort of authority or higher security this should prompt our theory to capture self-beliefs as having corresponding features. For instance, beliefs about one's own mental states might be more reliable than beliefs about the external world or other's minds. Moreover, one might think that because the question 'How do you know?' is inappropriate as a response to avowals beliefs about one's mental states have to be formed differently than any other type of belief. Given that 'What is your evidence?' or 'What is your judgment based on?' are also infelicitous responses to an avowal the methodology of the straightforward might lead one to propose that self-knowledge has to be non-evidential. Jordi Fernández describes this approach:

> […] one could appeal to some epistemic features of those beliefs as the reason why their expressions are basic and authoritative. As a matter of fact, it seems reasonable to think that an explanation of why our epistemic access to our mental states is special and strong could, in this way, deliver an explanation of why our assertions about them are basic and authoritative (Fernández, 2013, p. 10).

However, as indicated in the short discussion on norms of action, the linguistic practice underdetermines the epistemic practice. There might be similar problems related to a transition from our linguistic practice to features of self-knowledge. For instance, a reason

[14] I do not take a side in this debate and I am fully aware that I do not present it in a way that would fully capture it. The point is merely to illustrate a transition from language to epistemic matters and for that purpose the nutshell version of the debate should be sufficient.

why 'How do you know?' is inappropriate could be that the evidential basis is subpersonal, non-conscious and our inference from the basis proceeds very quickly. So it is not that self-beliefs are immediate or non-inferential and therefore we cannot answer the question, but rather we often cannot answer it because of some other fact about our cognition. Similar lines of reasoning have been proposed by Carruthers (2010; 2011) and Cassam (2014; 2017). The features of our avowals can be explained by more than a single set of features of our mental states and the workings of our brain. The straightforward way identifying single features of avowals with corresponding features of self-beliefs is an option, but not the only one. However, this criticism does not attack the link from observing knowledge ascriptions to learning things about self-knowledge in general. It only holds it to be an unreliable path that should be walked with caution.

Accounts accepting the doxastic view come by and large in one of two (nonexclusive) forms. One proposes that the features of avowals should be understood as based on a peculiar cognitive access to our mental states, the other holds that our linguistic practice is an indication of a peculiar relation of us as rational agents to our mental states. I discuss these in turn.

1.4.2 Cognitive Access

Proponents of this position (e.g. Armstrong (1968), Rosenthal (1986), Lycan (1987; 1996), Nichols and Stich (2003), Byrne (2005), Goldman (2006), Carruthers (2011), Fernández (2013), Cassam (2014)[15]) hold that avowals are reports of apparently *peculiar* and *privileged* beliefs. Hence, apparently peculiar and privileged beliefs are the explanandum we are interested in when we discuss self-knowledge. Here privileged means roughly, that beliefs about one's mental states acquired through introspection are more likely to amount to knowledge than beliefs about the external world or other people's mental states. (Byrne, 2005, p. 80) This is not necessarily to accept that one's beliefs about one's mental states are infallible. The claim is rather that it is easier to be correct about one's own mental states than about other people's states. However, an infallibility claim is compatible with a privileged access view, even though rarely defended nowadays. The reasons against infallibility rely on two pillars. First, infallibility is challenged by empirical data such as studies on the unreliability of verbal reports shown by Nisbett and Wilson (1977) and the studies of split-brain patients analyzed by Gazzaniga (1995); and second, it is further

[15] This list is not exhaustive.

attacked with the use of thought experiments such as the classic administrator case presented by Peacocke (1998):

> Someone may judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (Peacocke, 1998, p. 90).

If we take behavior to be a reliable indicator for beliefs it seems that someone can judge herself to have a certain belief, while actually having a different one. If this is correct, then we can be wrong about our own mental states. However, both the empirical studies and the thought experiment are challenged. Wilson (2002) himself argues against the significance of his earlier empirical studies, and Parent (2016) provides further rebuttals to empirical cases. Furthermore, Parent (2007) and Burge (1988) defend a limited infallibilism based on compositionality principles of thoughts, but they represent a clear minority in the debates around self-knowledge

Cognitive access accounts understand the peculiar nature of beliefs about one's mental states as these beliefs being formed by a peculiar method or way of knowing. Introspection is a process that differs from the methods that we use to acquire knowledge about the external world. Moreover, it is distinctly first-personal, such that it cannot be used to acquire knowledge of other people's mental states (Byrne, 2005, p. 81). Sometimes the peculiar nature is understood as beliefs about one's mental states being immediate or direct. This can be either in a psychological or an epistemological sense. The epistemological sense concerns justification. A belief is epistemically direct or immediate if its justification is non-inferential, that is, the justification for the belief does not come from other beliefs (not even in part) (Cassam, 2011; 2017). Notice that this only concerns the relations of two beliefs qua justification. A belief can be epistemically direct even though it requires another belief as an enabling factor (Pryor, 2005). A belief is psychologically direct or immediate only if it is not acquired by conscious reasoning or inference (Cassam, 2011; 2017). A belief can be epistemologically indirect, even though it is psychologically direct.

I formulated the nature of beliefs as apparently peculiar and privileged. The 'apparently' here is required to accommodate the 'non-straightforward' explanation of avowals. Carruthers (2010; 2011) and Cassam (2014; 2017) accept that avowals appear to be reports of beliefs that are formed in a peculiar and privileged way, but they reject that all (or even

most) avowals are actually reports of peculiarly formed beliefs. They hold that at least beliefs about propositional attitudes are not formed in a significantly different way than beliefs about other people's mental states. There might be some differences (such as the role of inner speech for Carruthers), but these differences are insufficient for self-beliefs about propositional attitudes to be special. The reason self-beliefs nevertheless appear to be special is that they are psychologically immediate, because they are by and large formed quickly and unconsciously.

Apparent peculiarity and privilege can come apart. For instance, one can try to understand our linguistic practice of avowing solely in virtue of peculiar belief, without privileged (or perhaps even underprivileged) access.[16] However, this seems to be a position difficult to defend given that we treat avowals as correct as long as we have no good reason for doubt.

1.4.3 Agentialism

Agentialist[17] positions (e.g. Burge (1996), Moran (2001), Bilgrami (2006)[18]) hold that what is special about self-knowledge is the relation of us as agents to our mental states. Even though not all agentialist authors explicitly discuss the relation to avowals, this relation can plausibly be the basis for the distinctive linguistic practice of avowing. What gives me authority in my avowals is the fact that it is *me* that I am talking about. I am the one responsible for my mental states. Believing or intending are things that I do, and not merely happen to me. Richard Moran formulates this idea:

> The phenomena of self-knowledge […] are themselves based as much in asymmetries of responsibility and commitment as they are in difference in capacities or in cognitive access (Moran, 2001, p. 64).

The agentialist thereby conceives the explanandum for self-knowledge as the special relation that one has as a (rational) agent to one's mental states and self-ascriptions of these states. Their views usually focus on attitudes and accept a cognitive access story for phenomenal states. Moreover, they focus on one's responsibility as a reasoner with mental states. Self-knowledge is taken to be an important tool to evaluate one's own mental states. One needs to know what mental state one is in to make sure that one can assess whether these mental states are properly formed, given the reasons one has. This can be used to form a transcendental argument for our ability to acquire self-knowledge, as

[16] Schwitzgebel (2008) argues that there is at least less privilege than we ordinarily assume.
[17] These are sometimes also called rationalist positions. (Gertler, 2011a)
[18] This list is not exhaustive.

provided by Burge (1996) and Moran (2003). Given that we can evaluate and change at least some of our mental states, and that this requires that we can know what mental states we are in, we can conclude that we actually are able to know our own mental states.

The possibility of changing some of one's mental states is crucial for the agentialist position. One can shape one's own attitudes – most importantly beliefs – by mobilizing and engaging with reasons (Moran, 2001). If one does so, one does not merely find out what attitude one has, but rather makes it the case that one has a certain attitude. For instance, if I think about whether Vienna is bigger in size than Edinburgh I can attend to relevant reasons and settle the question whether I believe that Vienna is bigger. And because I settled the question[19], my avowal that I believe Vienna is bigger than Edinburgh will be authoritative. The question for the agentialist is then how exactly this process works and what epistemic status the final verdict has. It is at least not obvious whether this deliberation results in true beliefs, knowledge, or something else (cf. Gertler (2011a, Chapter 6), O'Brien (2003)).

Given that the agentialist picture only addresses self-knowledge for attitudes it is an unlikely candidate for my overall project to provide a unified account of self-knowledge. Hence, this short description of general agentialist views will be sufficient to point to a different way of understanding the explanandum 'self-knowledge.' However, I will discuss Moran's view as a transparency account in more detail in Chapter 3, where I show why Moran's view ultimately even fails to generalize to propositional attitudes other than belief.

1.4.4 Explaining Self-Knowledge

If the explanandum is set up on the level of belief possible explanations are thereby restricted to the very same level. Hence, the apparently peculiar and privileged nature of beliefs about one's mental states need to be explained with reference to the way these beliefs are formed, and how this formation relates to the mental states the beliefs are about. This can be done in two explanatory frameworks: An agential-epistemological framework, and a psychological framework. The former is interested in the agential and epistemic features of self-beliefs and their formation, whereas the latter focusses on the cognitive process generating beliefs about one's mental states. On the agential-epistemological framework the explanation might involve especially good justification due to a peculiar belief-forming process, or a special relation between reasons and the mental

[19] For a discussion of 'settling a question' with respect to Moran see Vierkant (2015).

state in question that leads to well-formed second-order beliefs about these mental states. On the psychological framework the explanation will involve a story of how self-beliefs are produced in the brain, and how the architecture of our brain relates to these beliefs about our mental states. Both the agential-epistemic explanation and the psychological explanation ought to fit together. However, not everyone discussing self-knowledge is interested in both. It is a legitimate move to put the burden of a psychological explanation wholly on psychology and cognitive science and limit the philosopher's job to the agential and epistemic side of things. Nevertheless, an account that integrates explanations in both frameworks will be more explanatory powerful. Moreover, such an account makes predictions that can be empirically tested. For this reason Nichols and Stich (2003), Goldman (2006), and Carruthers (2011) discuss how their accounts relate to findings in psychological experiments.

Importantly, the doxastic view rules out any explanation on a non-doxastic level. For instance, if self-knowledge is understood as peculiar and privileged belief it cannot be explained by features on the level of our linguistic practice. Therefore, the doxastic view is incompatible with expressivism, the idea that avowals are only expressing mental states but not reporting them, and the features of our avowals are based on the special nature of expression. It is also incompatible with Wright's (1998) 'default view,' which explains the features of our avowals as *constitutive principles* of our mental state attributions. Here the idea is that these features of our avowals are part of the conditions of identification of what a subject beliefs, desires, etc. Your authority in avowing your belief is constitutive for you having the belief, rather than a privileged relation to your mental state.

In general, accounts of self-knowledge are incompatible with the doxastic view if their explanation of self-knowledge requires anything beyond the doxastic level, wherein doxastic level is understood as belief and belief-formation.

1.5 Conclusion

In this chapter I provided the starting point for the inquiry of self-knowledge. I accepted our linguistic practice as an overall starting point. This allows for two approaches of defining the explanandum 'self-knowledge,' which determine how potential explanations of self-knowledge can look like. First, one might take the linguistic practice at face value and conceive of the features of self-knowledge solely on the level of linguistic practice. I called this the linguistic view. I showed three different, but related ways to do so within the works

of Crispin Wright, Dorit Bar-On, and David Finkelstein. They all share that they define the features of self-knowledge with reference to avowals and responses to avowals. The linguistic view describes the features of self-knowledge solely by reference to linguistic practice, and therefore can remain neutral between epistemic and non-epistemic explanations of these features. Hence it allows for a non-epistemic account of the security of avowals as in Bar-On (2004). Nevertheless, the linguistic view is also compatible with an epistemic account that explains our linguistic practice based on epistemic features on the level of belief.

Second, one might take the linguistic practice as evidence for features on the level of belief. I called this the doxastic view. The doxastic view treats avowals as reports of beliefs about one's mental states. Moreover, it infers from our ordinary practice of avowing specific features of our beliefs and belief-formation. For instance, the fact that challenging avowals seems inappropriate leads to the idea that introspective beliefs are especially well justified, formed via a peculiar method, or a sign that one is in a special agential relation to one's own mental states. The doxastic view limits the possible explanations to epistemic and agential views.

In the next chapter I am going to argue that even though the linguistic view has the advantage of being more neutral, the doxastic view is more likely to capture all our folk intuitions about self-knowledge.

2 Beliefs over Avowals: Setting up the Discourse on Self-Knowledge

This chapter continues the discussion of defining the explanandum 'self-knowledge.' Starting with the two options presented in chapter 1 I explore whether one should be preferred over the other. I begin with a short introduction before presenting the motivation behind the linguistic view in part 2. I show that the linguistic view is more neutral and hence allows for more potential explanations of the peculiar features of self-knowledge. I also discuss how problems of epistemic accounts might be used as an argument in favor of the linguistic view. In part 3, I provide an original argument against choosing the linguistic view. Even though the linguistic view has advantages, it does not fully capture our folk notion of self-knowledge. I argue that their position lacks the tools to describe the role of self-knowledge in determining what one ought to do if someone disagrees with one's self-ascription of a mental state. If the argument is sound we have a conclusive reason to avoid the linguistic view. In part 4, I consider three ultimately unsuccessful attempts to respond to my argument. I finally conclude that our conception of self-knowledge has to be set up on the level of belief and belief formation to fully capture the phenomenon.

2.1 Introduction

In chapter 1 I discussed our linguistic practice as a starting point for capturing self-knowledge. Speech acts that self-ascribe mental states are a neutral, common ground to define what exactly the 'self-knowledge' we want to explain is. I introduced two broad ways to build on this starting point: the linguistic view and the doxastic view. Both are ways to capture the explanandum self-knowledge. However, they do so in different ways. The linguistic view identifies the explanandum directly with our linguistic practice; the doxastic view identifies it with features of belief and belief-production that are inferred from, or serve as explanation for our linguistic practice. The two options are captured in these principles:

(Linguistic View) The peculiar nature of self-knowledge should be completely described by features of linguistic practice, syntax, semantics, and pragmatics.

(Doxastic View) The peculiar nature of self-knowledge should be completely described by features of beliefs and belief formation.

Both positions presuppose that their suggestion can actually be followed through. We therefore have the corresponding presuppositions in place:

(Linguistic Presupposition) The peculiar nature of self-knowledge can be completely described by features of linguistic practice, syntax, semantics, and pragmatics.

(Doxastic Presupposition) The peculiar nature of self-knowledge can be completely described by features of beliefs and belief formation.

Both the linguistic view and the doxastic view aim to capture our folk notion of self-knowledge. According to the linguistic view the way in which self-knowledge is special and privileged is to be spelled out on the level of speech acts without missing out on any feature we would pre-theoretically attribute to self-knowledge. To do so a proponent to the linguistic view might identify the explanandum with a set of features of avowals. The doxastic view on the other hand characterizes the intuitive features of self-knowledge as results of a special and privileged belief-formation. My aim in this chapter is to discuss which of the two views on the table should be the basis on which an explanation of self-knowledge is built on. To do so I will first motivate the linguistic view, before providing an argument against it. I argue that even though the linguistic view has advantages, it fails to capture our folk notion of self-knowledge properly.

2.2 Motivating the Linguistic View

In chapter 1 I provided an overview of different variations of the linguistic view on the table. I will not take a stance on which formulation is preferable. Instead I focus on the more general principle (Linguistic View). Why should we choose the linguistic view over the doxastic one? The primary motivation to do so is argued for by both Wright (1998; 2015) and Bar-On (2004) extensively: We should prefer the linguistic view over the doxastic view, because the latter skews the discourse in a way that rules out some promising explanations of self-knowledge. This should not be particularly surprising, because the initial motivation to choose our linguistic practice as a starting point already was motivated in a similar way. Defining the features to be explained in terms of linguistic practice is more neutral than picking out what 'self-knowledge' is by reference to features of mental states or belief-formation directly. Nevertheless, it is worth looking at Wright's and Bar-On's concrete arguments to keep as much neutrality in the definition of the explanandum 'self-knowledge.' These arguments come in two shapes. First, both argue in favor of the neutrality of the linguistic view. Second, Bar-On argues further that epistemic accounts fail to explain our starting point of linguistic practice properly.

2.2.1 Neutrality

Bar-On (2004) argues that starting with questions about beliefs and knowledge brings about a dangerous temptation to only see one possible answer: some kind of especially secure method of making judgments about our present states of minds. If we start looking at doxastic states we will be blind to non-epistemic explanations of self-knowledge. Moreover, we will be tempted to overly assimilate avowals to assertions, which in turn might bring about problems to adequately explain what is special about self-knowledge. If we adopt the doxastic view we are forced to an epistemic approach. With the linguistic view on the other hand "we begin with a relatively neutral set of observations about the status of avowals, and try to understand that status […]. We can then take up questions of self-knowledge with a more open mind" (Bar-On, 2004, p. 13).

Wright (2015) has a similar aim in mind. Responding to Snowdon (2012) he provides the following rationale for the linguistic view:

> To set the record straight, then: I did not mean, in the Whitehead lectures, to side with Wittgenstein's view of these matters. But I did want to set things up in a way that allows his view to be heard (Wright, 2015, p. 55).

Here Wittgenstein serves as a placeholder for non-epistemic accounts in general. Wright's charge against the doxastic view is that any account that is not purely epistemic cannot enter the discourse if we set it up according to the doxastic view. Take for instance Wright's 'default view' as one such option. This is the idea that the authority of avowals is a constitutive principle that is not the result of any epistemic relation. He characterizes this view as follows:

> [T]he authority standardly granted to a subject's own beliefs, or expressed avowals, about his intentional states is a *constitutive principle*: something which is not a consequence of the nature of those states, and an associated epistemologically privileged relation in which the subject stands to them, but enters primitively into the conditions of identification of what a subject believes, hopes and intends (Wright, 1989, p. 142).

Wright is correct that this is not an available option if one starts with the doxastic view. The doxastic view presupposes that the peculiar nature of self-knowledge can be completely described on the level of belief and belief-formation. However, the default view explicitly rejects this option by stating that authority is something that avowals have by default, just by being avowals. There is no further fact that brings about the authority. Authority is a feature of the logical grammar of our speech (Bar-On, 2004, p. 347). This proposal that the

authority of avowals is something primitive to the speech act has no place in the doxastic view, because it posits that the peculiar nature of self-knowledge can only be explained as a linguistic feature of avowals. Only if the explanandum is conceived in terms of our linguistic practice this feature can be primitive. Regardless of whether Wright's approach is successful, leaving the option on the table is prima facie a virtue of the linguistic view. Furthermore, this line of argument is strengthened by Wright's claim that the linguistic view provides more options, while supposedly not rejecting any options that are permissible under the doxastic view. The linguistic view does not entail that

(Language Only) the peculiar nature of self-knowledge can *only* be explained by features of linguistic practice, syntax, semantics, and pragmatics.

The linguistic view merely claims that self-knowledge can be described in terms of linguistic practice, not that this description is irreducible. Perhaps we can still explain the linguistic practice in terms of privileged belief formation. Hence, Wright states that if we set up the explanandum in terms of avowals, an "[…] explanation in terms of cognitive advantage is by no means thereby ruled out […]" (2015, p. 54). However, with the linguistic view other explanations, such as Wright's default view or neo-expressivist[20] solutions, become available. These proposed solutions are not available for the doxastic view, because they deny any privilege on a doxastic level. A similar line of thought can be found in Bar-On, who holds that our account of the distinctive security of avowals should ideally leave a non-deflationary view of self-knowledge open. A description of the peculiar features of self-knowledge in terms of our linguistic practice should be compatible with an account on the level of belief (Bar-On, 2004, p. 20). Based on her neo-expressivist account of avowals she even provides different ways in which a non-epistemic explanation of the features of our linguistic practice could be connected to privileged beliefs (Bar-On, 2004, Chapter 9). However, just like Wright, Bar-On thinks the best way to define the explanandum is to focus on avowals and their features, so that non-epistemic explanations of the peculiar nature of self-knowledge are available in the first place. What we should explain is self-knowledge as a phenomenon on the level of our linguistic practice – especially secure avowals. If it turns out that this explanation fits with privileged belief states this is a welcome result, but the explanandum itself does not necessarily require anything on the level of belief. The peculiar nature of self-knowledge has to be understood in a way that does not rule out a negative

[20] E.g. Finkelstein (2003), Bar-On (2004; 2011), Bar-On & Sias (2013)

answer to the question whether one has privileged beliefs about one's own mental states. Bar-On herself does not give this negative answer (2004, p. 24), but her description of the especially secure nature of avowals is compatible with it.

The commitments of the linguistic view are supposedly small, which keeps the linguistic view more neutral than the doxastic view. Nevertheless, Proponents of the linguistic view are committed to (Linguistic Presupposition): the possibility of a complete description of the peculiar nature of self-knowledge by features of linguistic practice, syntax, semantics, and pragmatics. Furthermore, from this presupposition follows that any explanatory role that self-knowledge plays can be described by reference to features of linguistic practice, syntax, semantics, and pragmatics. Call this principle (Linguistic Features):

> **(Linguistic Features)** Any explanatory role that self-knowledge plays can be described by reference to features of linguistic practice, syntax, semantics, and pragmatics.

(Linguistic Features) is satisfied by all examples for the linguistic view in chapter 1 - Wright Bar-On and Finkelstein. They agree that self-knowledge talk can in principle be translated to talk about features of our linguistic practice. That I have privileged, or especially secure knowledge about myself can be translated to my avowals being protected from ordinary assessment, or the avowals being authoritative. If I say that I am in pain, my avowal will be accepted in normal circumstances.

2.2.2 Problems of Epistemic Accounts

Bar-On (2004, Chapter 4) provides another argument in favor of the linguistic view. In her discussions of different epistemic accounts of self-knowledge she provides a meta-inductive argument showing that epistemic accounts fall short of explaining the features of avowals. Even though the argument is not made explicit, the strategy is clear: Bar-On aims to show that the most promising epistemic accounts of self-knowledge cannot explain the security of avowals properly. If she is correct, then we have at least some reason to look for alternatives to epistemic explanations. These alternatives are only available if we step away from the doxastic view, and therefore we should think of the peculiar features of self-knowledge in terms of our linguistic practice. We should opt for the linguistic view.[21]

[21] A similar strategy is employed by Finkelstein (2003).

To establish the failings of epistemic accounts Bar-On quickly dismisses three different types of accounts of self-knowledge:

1. Cartesian versions of an infallible faculty that reveals one's own mental states to oneself.
2. *Material introspectionism*, which falls under peculiar cognitive access in the classification in chapter 1.
3. *Transparency accounts,* which can fall under peculiar cognitive access or agentialism according to the classification in chapter 1.

I limit myself here to her discussion of introspectionism. The infallibility condition alone is enough of a reason to rule Cartesian accounts. Moreover, I discuss transparency independently in chapter 3, where I also look into one of Bar-On's main attacks on transparency accounts[22] which states that they cannot capture self-knowledge for all types of mental states universally.

With regard to introspectionism Bar-On's discussion does not focus on a particular target, even though she refers to inner sense accounts of Armstrong (1968), Rosenthal (1986), and Lycan (1987; 1996). Her criticism of introspectionism is supposed to work against any view that proposes a distinct, especially reliable method of detecting one's own mental states by using a peculiar faculty of introspection (e.g. a mental scanner detecting one's beliefs). She opens a three-pronged attack on the inner sense assumption.[23] First, she argues that they have to accept brute errors of introspection. If we have a perception-like inner sense, then we should expect errors similar to the ones we make in visual perception. However, when confronted with an avowal that seems false, we do not treat it as simply an error of introspection, but rather take the avower to be at fault. Second, she claims that the introspectionist has to accept global failure of introspection, which does not fit our linguistic practice. She cites Wright (1998) as support here. In particular, she cites the very same passage I already noted in chapter 1 as especially puzzling to me:[24]

> You may not suppose me sincere and comprehending, yet chronically unreliable, about what I hope, believe, fear, and intend. Wholesale suspicion about my

[22] Bar-On's focus is on Evans (1982) and Moran (2001).
[23] Some of these arguments are not limited to inner sense accounts, but would, for instance, can also be used against self-other parity accounts of self-knowledge, such as Carruthers (2011).
[24] She also points in a footnote to Shoemaker's (1994) argument against self-blindness.

> attitudinal avowals—where it is not a doubt about sincerity or understanding—jars with conceiving of me as an intentional subject at all (Wright, 1998, p. 18).

Third, Bar-On proposes that the inner sense picture might work with phenomenal states (e.g. pain), but is a bad fit with intentional attitudes (e.g. beliefs or thoughts). Inner sense accounts have a difficult time explaining how they can detect intentional content embedded in attitudes such as beliefs.

I take these arguments not to be decisive against the introspectionist. An inner sense proponent can bite the bullet on the first and second argument, by either denying Bar-On's intuition on our practice of avowing, or by providing a further explanation for said practice. For instance, the fact that a false avowal is attributed to a failure of the avower could be a result of the low probability of brute errors in introspection. It is not that they are impossible, but the avower being insincere is simply much more likely, which explains why the go-to response to a false avowal is to blame the avower, and not assume a brute error. With regard to Bar-On's third argument the introspectionist could try to provide a convincing story of how content is represented in the brain such that knowledge of one's intentional states is possible. Nevertheless she provided prima facie reasons against some epistemic accounts on offer and thereby reasons to look for alternatives. Adding the advantage of being more neutral the linguistic view does seem to have something going in its favor. However, in the next section I am going to argue that the linguistic view comes with a drawback: it misses at least one feature we attribute to self-knowledge in our folk psychology.

2.3 The Argument against the Linguistic View

Having laid out the motivations for the linguistic view, my aim now is to challenge the view by targeting (Linguistic Features).

> **(Linguistic Features)** Any explanatory role that self-knowledge plays can be described by reference to features of linguistic practice, syntax, semantics, and pragmatics.

(Linguistic Features) is entailed by (Linguistic Presupposition) and therefore can be attributed to anyone accepting the linguistic view. My strategy is to find a feature of self-knowledge that cannot be spelled out in terms of linguistic practice. This undermines the linguistic setup by putting pressure on the idea that the linguistic view fully captures our

folk notion of self-knowledge. There might still be a hybrid view that accepts self-knowledge as having features on multiple levels, but setting up the problem solely on the level of linguistic practice will be hopeless. If the argument is successful it thereby threatens Wright's position and the neo-expressivist project by undermining their starting point. In a nutshell the argument works as follows:

> P1: The features of self-knowledge play a role in philosophical problem X, so they should tell us something about problem X.
>
> P2: No feature of our linguistic practice is going to tell us anything about philosophical problem X.
>
> C: Self-knowledge cannot be captured wholly in features of our linguistic practice.

Given (Linguistic Features) any explanatory role that self-knowledge plays can be described in terms of features of our linguistic practice, syntax, semantics, and pragmatics. Now suppose a case of a philosophical problem in which intuitively self-knowledge plays an explanatory role. Based on (Linguistic Features) we should be able to re-describe the role of self-knowledge in this case in terms of features of the linguistic practice. However, if there is no adequate description that substitutes the folk notion of self-knowledge with features of our linguistic practice, then it looks like self-knowledge cannot be captured completely by features of our linguistic practice. Hence (Linguistic Features) must be false. Moreover, because (Linguistic Features) is an entailment from (Linguistic Presupposition) we can conclude that (Linguistic Presupposition) is false, and the linguistic view collapses.

The problem in question is the case of a *mental state disagreement*[25]: A subject S believes (by non-interpretational means[26]) that S is in mental state M. An interlocutor L claims that S is not in M, but rather in a different state M*. We then ask, whether S ought to change her confidence in her belief in the face of disagreement. To illustrate mental state disagreements consider the following case:

[25] I do not claim that is the only problem that can be used for this line of argument.

[26] This clause is supposed to rule out cases in which I believe that I am in a mental state based on observing and interpreting my behavior. Some philosophers argue that there is no non-interpretational self-ascription of attitudes. They propose that self-ascriptions are based on interpretation, even though we might be in a better position to observe ourselves than others. Cf. Carruthers (2011), Cassam (2014)

(Friend) Suppose I am part of a university admission committee. I know that I am supposed to be fair towards all applicants. I explicitly state that I treat every submission the same, regardless of the ethnicity of an applicant. "I believe ethnicity makes no difference in the quality of a candidate," I say. However, my friend disagrees. She knows me well and she also knows all my 25 past decisions on the committee. We disagree about my mental state. Being confronted with this disagreement, what am I to do?

In this case it seems plausible to concede that I might be biased.[27] Perhaps I actually believe that people of my ethnicity are better applicants, but I lack awareness of this belief. Given that my friend knows all my past decisions, it seems that I should accept her testimony and lower the confidence in my belief.

We can further characterize mental state disagreements. First, I can be wrong in assessing my mental states, and my interlocutor can be right. I do not argue for this here, but merely point out that it is accepted by proponents of the linguistic view (cf. Wright (1998; 2001; 2015), Bar-On (2004), and Finkelstein (2003)).

Second, interlocutors in mental state disagreements are not epistemically equal. My friend and I form our beliefs on a different basis. She observes my behavior and infers my belief, I introspect. We are also aware that we form our beliefs differently and are not epistemic peers. Moreover, interlocutors in mental state disagreements do not only happen to be epistemically unequal, they are in principle unequal. They cannot get on the same epistemic level by disclosing their evidence. This inability is one-sided. My friend has no problem to disclose her evidence. She can state what behavior she observed. However, I am unable to disclose my evidence. Whatever the basis for my belief is (if there is one), I cannot access it. My only option appears to be stating that 'I just know.' Full disclosure is ruled out. Hence, there is no way to get on an equal epistemic footing.

One might suggest that this is not a problem. Even though full disclosure is ruled out in the sense that I cannot share my evidence, it should be enough that my interlocutor can cite her evidence. I can then adjust my belief on the total evidence, mine and hers. However, it is unclear how exactly this is supposed to work. I cannot weight her evidence against mine, because that would also require access to my evidence for my mental state.

[27] This presupposes that actions are a guide to belief. I take this to be a plausible assumption, although not completely uncontroversial.

Third, there is no universally ideal response to all mental state disagreements.[28] In (Friend) I ought to change my confidence in my second-order belief. However, consider a slightly different case:

> **(Passer-by)** Suppose I am part of a university admission committee. I know that I am supposed to be fair towards all applicants. I explicitly state that I treat every submission the same, regardless of the ethnicity of an applicant. "I believe ethnicity makes no difference in the quality of a candidate," I say. A passer-by disagrees. She does not know me well, and she only knows a single decision of mine on the committee. Being confronted with this disagreement, what am I to do?

It seems out of the question to revise my belief here. My interlocutor barely knows me, and she has little evidence for her judgment. I ought to stick to my belief and confidence level. Because some cases require one to lower one's confidence, while other cases rationally require one to hold onto one's confidence level there clearly is no universal response to mental state disagreements. The difference between (Friend) and (Passer-by) seems to be the epistemic standing of my interlocutor. The more evidence my interlocutor can cite, the more rational it seems to decrease confidence in my belief. This is not surprising as it is a feature of disagreements in general. The more justified I take a disagreeing interlocutor to be, the more rational it is to adjust my belief based on the disagreement. Furthermore, I have to believe that my interlocutor surpasses some threshold of justification before the disagreement rationally requires me to decrease confidence in my belief at all. As long as I do not take the interlocutor to pass this mark, the disagreement is ineffective.

In mental state disagreements the threshold for rationally required change in confidence seems to be higher than in ordinary disagreements. This can be illustrated by considering two parallel cases, one involving disagreement about a third person's mental state, and one involving disagreement about my mental state.

a) Suppose I am at a party. Kate, a friend of mine, and John, my long-time colleague are also present, but they are in a different room at $time_1$. Later at $time_2$ I talk to Kate about 80s music. I sincerely say that John believes 80s music is terrible. Kate,

[28] I aim to keep my commitments to views in the debates on the epistemology of disagreement as minimal as possible. However, I am committed to the idea that not every disagreement demands the same rational response. Some disagreements require one to be conciliatory; others require one to be steadfast. One view in the literature that allows for both is Lackey's (2008; 2010) *Justificationist View*.

who observed John picking out music at the party at $time_1$, disagrees. She thinks John actually believes that 80s music is good, and it shows in his behavior selecting typical music from the 80s at the party. However, Kate only met John 3 months ago, and hence does not know John very well. Should I lower the confidence in my belief that John believes 80s music is terrible?

b) Suppose I am at a party, choosing some music to play. Kate, a friend of mine, is also present. Later I talk to Kate about 80s music. I sincerely say that I believe 80s music is terrible. Kate, who observed me picking out music at the party earlier, disagrees. She thinks I actually believes that 80s music is good, and it shows in my behavior selecting typical music from the 80s at the party. However, Kate only met me 3 months ago, and hence does not know me very well. Should I lower the confidence in my belief that I believe 80s music is terrible?

I suggest that the answer to (a) is yes, and the answer to (b) is no. In the latter case I know myself better than my friend knows me. I know myself better to a degree that makes it permissible to disregard her disagreement. However, I cannot disregard her testimony about my colleague in case (a), because I do not know my colleague in a similar way. The disagreement about someone else's mental state requires less to be effective than the disagreement about my own mental state. If this is correct, then it seems that the involvement of self-knowledge makes a difference for determining the rational response to a disagreement. This is the crucial step in the argument against the linguistic view. Once you accept that intuitively self-knowledge plays a role in determining the rational response to mental state disagreements the linguistic view is in trouble.

Given that self-knowledge plays this role in mental state disagreements we can search for ways in which self-knowledge might play this role. The obvious choices to look into are justification and confidence. If self-knowledge is thought of as a product of peculiar belief formation that provides especially strong justification, then the influence on a rational response to disagreements is straightforward. I am more justified in my mental state self-ascriptions than I am in attributing mental states to others. Hence, I can stick to my belief in (b), but cannot do so in (a). This is a perfectly fine explanation for advocates of the doxastic view. However, for proponents of the linguistic view the answer cannot be justification or confidence. They are committed to (Linguistic Features), the claim that the role that self-knowledge plays can be described by reference to features of the linguistic practice, and

neither justification, nor confidence[29] in beliefs seem to be part of the linguistic practice. The challenge is that they need to find a way in which my rational response is influenced by self-knowledge without relying on any epistemic differences between (a) and (b). This challenge seems impossible to meet. Any characterization of self-knowledge in terms of linguistic practice will be quiet on how I ought to rationally change my beliefs in any situation. After all, rational beliefs are not part of the linguistic practice.

The problem is that the linguistic view defines the features of self-knowledge in terms of our linguistic practice with reference to appropriate or inappropriate speech acts, either avowals, or responses to avowals. However, in a mental state disagreement we already start with an avowal and a response. Whether they are appropriate does not seem to matter at this point. All we are interested in is what to do *after* the initial response by my interlocutor. All linguistic features are already out of the game once the disagreement enters. Nevertheless, self-knowledge seems to play a role here, because our rational response can differ between mental state disagreements and corresponding ordinary disagreements.

Take a second comparison. This time let us compare a disagreement case involving an interpretation based self-ascription with a disagreement involving genuine self-knowledge.

c) Suppose Anna and I are candidates for a job. Anna gets the job, while I have to keep looking for work. In the next days I notice that I act a little hostile towards Anna. My parts in our conversations are short and my tone is rather unfriendly. Looking at my own behavior I conclude that I must be envious. Talking to Anna I apologize and tell her that I'm envious which is why I act so rude. She disagrees, telling me that I'm just frustrated that I have to keep looking for a job, I'm not really envious. Anna knows me well. Moreover, I know that Anna is always blunt and is sincere in her assertion. Should I lower the confidence in my belief that I'm envious?
d) Suppose Anna and I are candidates for a job. Anna gets the job, while I have to keep looking for work. In the next days I act a little hostile towards Anna. My parts in our conversations are short and my tone is rather unfriendly. Without noticing this behavior I believe that I'm envious of Anna. Talking to her I apologize and tell her

[29] Remember that even though Bar-On (2004) mentions a high degree of confidence of one's avowal this cannot be an epistemic notion of confidence if she wants to hold on to the linguistic view. Hence she cannot use confidence in beliefs to explain the difference between (a) and (b).

that I'm envious. She disagrees, telling me that I'm just frustrated that I have to keep looking for a job, I'm not really envious. Anna knows me well. Moreover, I know that Anna is always blunt and is sincere in her assertion. Should I lower the confidence in my belief that I'm envious?

I suggest that just like in (a) and (b), the answer to (c) is yes, and the answer to (d) is no. One might suspect that the linguistic view has the tools to deal with this case, because its proponents make a distinction between proper avowals and assertions based on self-interpretation. However, the linguistic view still lacks the tools to tell us why lowering the confidence in my belief would be *rational* in one case, but not in the other. All the linguistic view can tell us here is that in (c) Anna's disagreement is appropriate, but in (d) it is not. Once again, the linguistic view does not tell us anything about our response after Anna actually disagreed, but self-knowledge still plays a role at this point. The linguistic view fails to fully capture our folk notion of self-knowledge.

2.4 Against the Argument

Proponents of the linguistic view cannot respond by simply denying that mental state disagreements exist. They definitely do, and moreover they seem unavoidable if one accepts fallibility for self-ascriptions. However, I think there are at least three different, interesting ways to respond to the argument for the proponents of the linguistic view. First, they may object that I underdescribe the cases and hence we cannot be sure what our intuitions should be here. Second, they may argue that I smuggled the doxastic view into my premises when I set up the problem as a question about a rational response to a mental state disagreement related to my confidence in my belief. Third, they may deny that genuine self-knowledge plays a role in these disagreements.

2.4.1 The Case is Underdescribed

Friends of the linguistic view can disagree with my suggested answers for cases (a) to (d). Furthermore, they might contest that it is generally unclear what to think about these cases, because they are not sufficiently well described. In (a) and (b) we don't know exactly how well people know each other. John is a long-time colleague, but what exactly does that mean for my knowledge of John's typical behavior? Perhaps I know John so well that I can safely ignore Kate's disagreement. Or perhaps Kate knows neither John nor me well enough for her disagreement to matter. Similarly, in (c) and (d) the cases do not state how well exactly Anna and I know each other, nor what my behavior in the case is exactly.

However, I do not think that any charitable way of filling in more details is going to change the intuitions pumped. All that is required for the argument is that we compare a case in which the belief was intuitively formed by observation, inference and interpretation, with a case that is intuitively an instance of self-knowledge (or at least self-belief). The latter has to be stated such that it is easily recognized as an instance that does not appear to be based on observation, inference and interpretation. As long as we hold this difference in intuitive belief-formations fixed we can add as many details as we want. For instance, I can add that Anna and I know each other for three years and we meet about once a week to (c) and/or (d). I might further add that we usually have long conversations, especially about topics we care about. Perhaps, we talk about a topic I usually am enthusiastic about, but nevertheless I clearly attempt to end the conversation as quickly as possible. These additions are not changing the intuitions pumped, even in case they introduce differences between (c) and (d). The intuition pumped is still that I seem to require different adjustments to my beliefs in (c) compared to (d). Hence, the charge of underdescription seems to miss the point.[30]

2.4.2 Begging the Question

The second objection raises an issue about the setup of my argument. I ask what is rational to do in mental state disagreements. Should I lower the confidence in my belief or should I stay put? However, the proponent of the linguistic view might object that this already locates the discussion to the level of belief. They would describe mental state disagreements differently: A subject S avows that S is in mental state M. An interlocutor L claims that S is not in M, but rather in a different state M*. We then ask, whether it is appropriate for S to avow that S is in mental state M in the face of disagreement. Here no talk of beliefs or rational response is present.

The immediate response here is to ask whether this is the right way to describe mental state disagreements. In ordinary disagreements it is fine to ask the question of what one is supposed to believe in the face of disagreement, so why should it not be equally fine to ask the question when the disagreement is about my mental states? Moreover, based on our ordinary linguistic practice there does not seem to be anything wrong with wondering what

[30] Thank you to an anonymous reviewer for Episteme pointing out that my response to this point in an earlier version was misdirected, and that a straightforward answer allowing details to be added in any charitable way is available.

one ought to believe in a mental state disagreement. So this objection has to be further motivated to get off the ground.

Let us suppose it can be sufficiently motivated. Even in this case, we can run a version of my overall argument, because in some disagreements of this kind it seems appropriate for S to still avow that S is in mental state M, while in others it is not. And the difference between these cases has to be accounted for without any epistemic difference, which seems to be challenging.[31]

However, perhaps even this description of mental state disagreements is not acceptable for the linguistic view philosophers. They could argue that the interlocutor L already acted inappropriately by voicing her disagreement as a response to the avowal. Furthermore, the question how one appropriately responds to an inappropriate speech act is misguided. No inappropriate speech acts demands a particular response. One cannot be blamed for any response to an inappropriate speech act. There is no rule in our linguistic practice to govern inappropriate challenges of my avowal. Once my opponent stops playing by the rules I cannot look at the rules for what to do. Just as there is no legitimate chess move as a response to someone stacking the Knight on top of the Rook, there is no proper move in the language game after a mental state disagreement.

The problem with this response is that mental state disagreements do not look like a complete breakdown of the rules of communication. Mental state disagreements can be appropriate, even though they are rare occurrences. Moreover, when someone disagrees about my mental state, there appears to be a right and a wrong response for the particular case. We have not stopped playing our language game. I take this to be an indication that our folk notion of self-knowledge includes the possibility of challenges by others. There is a right response to such challenges, and our theoretical conception of self-knowledge should provide enough tools to fully explain why we ought to respond a certain way. The linguistic view cannot do that, and hence fails to capture our folk notion of self-knowledge.

2.4.3 No Genuine Self-Knowledge

Finally, friends of the linguistic view may deny that genuine self-knowledge plays a role in mental state disagreements. One can make the case that Bar-On (2004) has a response of

[31] Perhaps the proponent of the linguistic view could reiterate their initial response here and claim that this is just how our language works, no further explanation needed. So there might be a way out if they can show that their way of describing mental state disagreements is right.

this kind built into her set-up. Her conception of the explanandum for self-knowledge involves avowals being "[…] protected from ordinary epistemic assessments (including requests for reasons, challenges to their truth […] etc.)" (p. 20). This gives her the option to rule out mental state disagreements as cases of *extraordinary* epistemic assessments. Furthermore, she provides examples of extraordinary mental state self-ascriptions, including "[…] on the basis of therapy, *consultation with others*, self-interpretation, or cognitive test results" (p. 194, emphasis added). On this basis Bar-On can argue that the question of a rational response to mental state disagreements is not relevant for genuine self-knowledge, because the mere fact that I take the interlocutor's disagreement seriously indicates that I left the ordinary linguistic practice of avowing, and entered a different language game. One in which my statement looks like an avowal, but is treated as a mere report. In other words, Bar-On can contest whether the impact of self-knowledge in mental state disagreements is actually part of our folk notion of self-knowledge.

This response seems to presuppose a too narrow scope of self-knowledge. It seems arbitrary to posit that certain sincere and largely non-interpretative mental state ascriptions are not avowals, while others are. Bar-On might respond that it is not arbitrary for two reasons. First, because in mental state disagreements I consider evidence my interlocutor provides. My self-ascription after taking my interlocutor's disagreement seriously appears to be partially inferred from evidence. Hence, it will be different than ordinary avowals which appear to be non-evidential (Bar-On, 2004, p. 2). Second, all avowals show immunity to error through misascription. They do not involve any recognition of a mental state and therefore are protected from epistemic assessment (Bar-On, 2004, Chapter 6). My response to the mental state disagreement on the other hand is an attempt to recognize my mental state correctly with the help of my interlocutor's testimony, hence it is not an avowal.

However, both reasons can be challenged. First, even though the interlocutor's testimony constitutes evidence, the mental state ascription can still be largely non-evidential. I am not self-ascribing a belief *only* on the testimony. Plausibly I can have my own, non-evidential judgment, which I then adjust based on the testimony. The result appears neither fully evidential, nor fully non-evidential. The question is whether we should treat it like the fully

evidential, or like the fully non-evidential case.[32] I think there is a prima facie reason against the former. After I adjust my belief according to the disagreement I can still be the authority regarding my self-ascriptions. This authority is something that is not present in case I assert solely based on my evidence. There is still an asymmetry between the self-ascription in the post-testimony case and ascriptions of others' mental states (or ascriptions of my own mental states based fully on evidence). If this is correct, then we should treat the self-ascription after considering testimony more akin to genuine avowals.[33]

Second, arguing from the immunity to error through misascription gets the order of explanation wrong. The concept of immunity to error through misascription is introduced by Bar-On (2004, Chapter 6) to explain why avowals are protected from ordinary epistemic assessment. Avowals, so Bar-On, do not involve any recognition of a mental state. Moreover, you cannot accuse me of making an epistemic mistake, if I did not perform any epistemic action at all, so challenges to my avowal are off the table. However, given that this feature is supposed to explain a property of avowals we cannot use it to pick out which speech acts are avowals. We want to find out whether an explanation in terms of immunity to error through misascription fits the folk notion of self-knowledge, and therefore we should not pick out the extension of this folk concept in virtue of the theoretical concept in question.

Given that these reasons to disregard the responses to mental state disagreements as cases of genuine self-knowledge do not hold up well we are left without any principled way to rule out mental state disagreements as ordinary interactions involving self-knowledge. Hence, we are stuck to appealing to intuitions. And it does not seem intuitive to treat my claim 'I believe that p' differently as soon as someone disagrees and I take the disagreement seriously. My speech act does not appear to change once I wonder whether I should change my belief in the face of disagreement.

2.5 Conclusion

In this chapter I developed the motivations behind the linguistic view, before arguing that it ultimately fails, because it does not fully capture our folk notion of self-knowledge. To show

[32] One might raise here the possibility of treating it like neither of these options, but I do not know how the alternative would look like.

[33] One might object that Bar-On could treat the post-testimony case as being fully based on evidence, with the previous avowal being part of the evidential basis. However, this would still be incompatible with the authority that one has when one self-ascribes after considering the testimony.

this I argued that mental state disagreements are cases in which the nature of self-knowledge affects my rational response to the disagreement. Given this feature, any account of self-knowledge has to be able to explain how self-knowledge influences what I ought to do. I proposed that accounts starting with self-knowledge as fully describable by linguistic practice cannot do that. Therefore setting up the problem according to the linguistic view seems to be misguided. Instead, we ought to explain self-knowledge as a phenomenon on the level of belief and belief formation. If this is correct, then the neo-expressivist project with a sole focus on our linguistic practice should be abandoned and rethought as something starting from a hybrid view, instead of the linguistic view. For my own project this means that I opt for a characterization of self-knowledge based on the doxastic view: The peculiar features of self-knowledge are described by reference to features of beliefs and belief-formation. Beliefs about one's own mental states are usually[34] formed differently than beliefs about other's mental states. Moreover, beliefs about one's own mental states formed in this peculiar way are usually more reliable than beliefs about the external world or other people's mental states.

In chapter 3 I move from setting up the explanandum 'self-knowledge' to the explanation of the peculiar nature of self-knowledge. I motivate my choice to develop a transparency account of self-knowledge and discuss what exactly such an account should look like to overcome common problems of transparency accounts.

[34] In some cases one's own mental state might only be accessible via self-directed mind-reading. In these cases one observes one's own behavior and then infers one's own mental state.

3 Transparency

This chapter provides an introduction to the type of explanation I want to propose: *a transparency account of self-knowledge*. My aim here is to provide the conceptual landscape of transparency accounts and show to what extent vastly different views of self-knowledge have a common core that qualifies them as transparency accounts. This serves as motivation for my view and helps to avoid standard objections to transparency accounts. I start with a discussion of Evans (1982) and why we should be interested in transparency accounts in the first place. In part 2 I determine general features that define a transparency account, before making the distinction between 'move' and 'no-move' accounts in part 3. In part 4, I provide the taxonomy for move accounts, including case studies of Moran (2001) and Byrne (2005). Part 5 is a discussion of the corresponding taxonomy for no-move accounts, with a case study of Boyle (2009; 2011). In the remaining section 6 I look at the problem of scope and the standing state problem for transparency accounts. I relate them to the taxonomies and argue that these problems are closely connected to certain paths in the taxonomy. I suggest that we can solve them by avoiding these paths.

3.1 Introduction

In the last chapter I set up the explanandum 'self-knowledge' according to the doxastic view. This set up the agenda for the rest of the enquiry as follows: I want to explain why beliefs about one's own mental states are usually formed differently and more reliably than beliefs about other's mental states. This chapter provides an introduction to the type of explanation I propose: *a transparency*[35] *account of self-knowledge*. My aim in this chapter is to discuss the core idea that self-knowledge is transparent. However, vastly different accounts fall under the same label 'transparency.' Prima facie there seems to be little in common between Gallois's (1996) inferential approach and Boyle's (2011) metaphysical assumption that a belief and knowledge of that belief are a single psychological state. I provide an overview of the conceptual landscape of transparency accounts, showing how views of self-knowledge that seem so different on first sight have a common core that qualifies them as transparency accounts. I start with a discussion of Evans (1982) and motivate why we should be interested in transparency accounts in the first place in part 2. In part 3 I determine general features that define a transparency account, before making

[35] Note that there are different discussions surrounding the term 'transparency.' I am only concerned with Evans style transparency, which has to be distinguished from the debates on transparency/diaphanousness by G.E. Moore (1903) or Gilbert Harman (1990). Moreover, sometimes (e.g. by Carruthers (2011)) the term 'transparency' is used interchangeably to luminosity, the claim that one knows, or is in a position to know that one is in a mental state simply in virtue of having that mental state.

the distinction between 'move' and 'no-move' accounts in part 4. In part 5, I provide the taxonomy for move accounts, including case studies of Moran (2001) and Byrne (2005). Part 6 is a discussion of the corresponding taxonomy for no-move accounts, with a case study of Boyle (2009; 2011). In the remaining section 7 I look at the problem of scope and the standing state problem for transparency accounts. I relate them to the taxonomies and argue that these problems are closely connected to certain paths in the taxonomy. I suggest that we can solve them by avoiding these paths.

3.2 A First Look at Transparency

In the last twenty years we had no shortage of self-knowledge accounts that take themselves to spell out the idea of 'transparency' – the idea that self-ascribing mental states is done by attending outwards instead of inwards.[36] They all trace themselves back to Evans (1982)[37], with most of the work on transparent self-knowledge understood as ways to spell out Evans's (1982) remarks. The central passage in *The Varieties of Reference* is this:

> [I]n making a self-ascription of a belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p. (There is no question of my applying a procedure for determining beliefs *to something*, and hence no question of my possibly applying the procedure to the wrong thing.) If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined sceptic cannot find here a gap in which to insert his knife (Evans, 1982, p. 225).

Evans aims to provide an alternative to inner-sense views of self-knowledge, which he took to be the dominating account. The paradigmatic 20th century example for such an inner-sense picture is Armstrong's (1968) account of perception-like inner-sense. On this view we are equipped with a scanning faculty that simply detects our mental states reliably. Evans wants us to discard the idea of anything like such a scanner. He pumps the intuition that in general we do not look inwards in any sense when finding out what we believe, but rather attend to outward phenomena. We focus on whether p is true, when we want to find out whether we believe that p.

[36] For instance Gallois (1996), Moran (2001), Byrne (2005), Boyle (2009), Roessler (2013), Fernández (2013), Silins (2013), and Antonia Peacocke (2017).
[37] And to a lesser degree to Edgley (1969).

Evans provides no direct argument in favor of this procedure. However, there are at least three reasons in favour of an inquiry of transparency. First, as the prime motivation for Evans, we look for an alternative to the picture of gazing inwards. Transparency is not the only candidate for such an alternative, but it is certainly a promising one considering that – second – transparency ought to be ontologically *economic*. That is, we want transparency accounts to be undemanding in terms of additional entities or processes in contrast to theories that require newly posited processes or faculties (such as Armstrong's scanner). The feature of being economic is already present in Evans (1982, pp. 225-226) and further emphasized by current transparency theorists such as Alex Byrne (2005; 2011; 2018). Third, as part of accounts of self-knowledge transparency is supposed to be partially explaining the key features of self-knowledge. The asymmetry between self-knowledge and knowledge of other's mental states and the high reliability of self-ascriptions ought to be explained with the help of transparency. If a more precise grasp on transparency is helpful to do all that it is certainly a valuable enterprise.

There is an additional, though more contentious reason in Evans's discussion: an appeal to our phenomenology. Evans introduces the idea by providing an example of what one does when being asked 'Do you think there is going to be a third world war?' The reader is supposed to be convinced that Evans's description of what follows in forming a response to the question is intuitively correct. So if someone were to ask the reader of this passage the very same question, they would form an answer in the very same way. It might be a different answer, but it would still be formed by attending to whether there will be a third world war. Evans's intuition pump seems to be effective to some degree, given that a significant number of philosophers took this as a fruitful starting point for building theories of self-knowledge. However, there seem to be countless examples in which the phenomenology of forming beliefs about one's mental states does not seem to fit the same story. Suppose you were asked whether you believe that we are in the year 2018. Do you experience yourself as attending to reasons for being in the year 2018? – Probably not. You can answer such a question immediately, without any awareness of attending to anything at all. Your second-order belief seems to be just there, ready for you to access. This is not a phenomenon that can be explained away by focussing on the observation that we sometimes use questions whether one believes that p as substitutions for questions whether p. Consider a similar question about desires: do you desire a coffee? Again, it

seems unquestionable that sometimes[38] you can answer immediately without any experience of attending to the coffee. In O'Shaughnessy's (2000) terminology: self-knowledge seems to be formed silently. Usually we are not aware of how we know our own mental states. Even if there are cases in which the phenomenology supports the transparency proposal, there are other cases which do not. Hence, the phenomenology is not going to be a good motivation for a transparency account. Nevertheless, given that a transparency account might still provide an economic and powerful explanation this should not deter anyone from exploring transparency accounts. However, it does provide another desideratum for an ideal transparency theory insofar as such a theory should give us some idea why our phenomenology of self-knowledge is the way it is.

3.3 Defining Transparency

I sketched reasons why we should think about transparency for self-knowledge. But what exactly does transparency amount to? Evans starts with the idea of an outward direction, which sets up the position as an alternative to inner-sense accounts. Moreover, he uses the notion of an outward phenomenon as that which one attends to for self-ascriptions. Evans is quiet on what exactly the outward phenomenon is. He gives an elaborate example in the third world war case, in which, intuitively, one weights and compares reasons for and against another world war happening. This seems to be a process of conscious deliberation. So one might be lead to think attending to the outward phenomenon has to be such deliberation. However, there is no evidence in his writings that he thought conscious deliberation based on reasons would be necessary. What he is after is the idea that I attend "[…] to precisely the same outward phenomena […]" (Evans, 1982, p. 225). That is, whatever I attend to for first-order belief-forming is also the same thing I attend to for self-ascribing first-order beliefs. He tries to remain neutral on what exactly the outward phenomenon is. Of course, he has to indicate one specific outward phenomenon insofar as he chooses an example to illustrate the idea, but that is not what he wants to get at. It is the identity of epistemic basis for first-order beliefs and their self-ascription that he is aiming for.

This is further emphasized when he talks about sameness of procedure. He states, that "I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (Evans, 1982,

[38] There might be cases in which you are unsure whether you want a coffee.

p. 225). Once again there is no claim about the procedure in question. Moreover, the general phrasing of the statement is also evidence against the necessity of any particular reading of 'attending to an outward phenomenon.' Evans proposes a sameness of procedure, which implies that whatever procedure is used to form a first-order belief is also used to form a second-order belief. In the third world war case the first-order belief is formed by pondering on reasons for and against another world war happening. Hence, the self-ascription has to be done in virtue of the same method – deliberating on those reasons. However, because Evans states his sameness of procedure in method-neutral terms, it is easy to see that the procedure can come apart from such deliberation. If one forms a first-order belief by perceiving the world one has to self-ascribe this belief in the same manner, lacking conscious deliberation completely.

The point to take away is that Evans's transparency proposal is of a general kind. It is defined by sameness of outward phenomenon, and sameness of procedure with which we attend to that outward phenomenon. There is no specific way to spell out either phenomenon or procedure in Evans. This leaves the question of what transparency exactly is wide open. However, there are structural conditions that have to be satisfied. Any way to spell out transparency has to have an outward phenomenon as a starting point. Moreover, it also has to involve some way of self-ascribing a mental state based on this outward phenomenon. This is the way of spelling out 'attending to an outward phenomenon.' Attending thereby is not conscious awareness of the outward phenomenon. Rather, it is the procedure that takes one from the outward phenomenon to the self-ascription. This procedure has to be the same as the first-order belief-forming procedure.

Two remarks about 'sameness' of procedures: First, I want to emphasize the idea of same procedures. Discussions of transparency frequently involve talk of a 'transparency method'[39] or 'transparency procedure' that might mislead one to postulate a special method of transparency. Consider Bar-On's otherwise careful discussion of transparency in which she describes Evans's transparency account as follows: "By invoking a special transparency procedure or method, then, Evans can capture the contrast between intentional avowals and other intentional ascriptions" (2004, p. 110). This formulation makes it seem as if we had a distinct, special method of transparency. However, this is in conflict with Evans himself. He clearly states that I use the same procedure as in answering

[39] E.g. Moran (2001), Bar-On (2004), Cassam (2014)

the question whether p. I do not need a special method at all. I use the most normal method there is. This is important for one of the core motivations to explore transparency: the economy of an explanation. If the transparency method was special, it would not be economic, because it would propose peculiar, distinctive ways of introspection.[40] To Bar-On's credit, her further exposition of Evans's thought explains that we engage "[…] the ordinary abilities and dispositions that result in our acquisition of first-order intentional attitudes about states and objects in the world outside us," (2004, p. 110) which captures Evans's idea properly. However, her discussion provides a good example of how easy certain terminology can lead us astray. Hence, any definition of transparency ought to emphasize the sameness of procedures.

Second, 'sameness' here is a puzzling notion. On one hand, if it were exactly the same, then self-ascription and first-order belief formation cannot come apart at all. So if you form a first-order belief, you also form a self-ascription of that belief. And there is some indication that Evans intends to talk about 'exactly the same procedure.' After all he proposes that one "[…] goes through exactly the same procedure as he would go through if he were trying to make a judgement about how it is at his place now […]" (1982, p. 227) when talking about transparency related to perception. On the other hand, Evans also uses the notion of "[…] re-using precisely those skills of conceptualization that he uses to make judgements about the world" (1982, p. 227), and 're-using' would not make any sense if it were exactly the same procedure. What would re-using add to the picture if it would be an iteration of exactly the same procedure?

I take it that Evans work is inconclusive regarding how strict we should read the notion of 'same procedure.' A *strong reading* has it that it is exactly the same procedure and re-using does not provide anything new (but perhaps has some other use). A *weaker reading* only requires that there is no distinct method of self-ascribing. The idea is that self-ascriptions of beliefs are latched onto first-order belief-forming processes. They are generated by the same procedure because they cannot occur independently of first-order belief formation. Finally, the *weakest reading* is a claim that the method used in forming beliefs about one's mental states is not limited to self-ascribing mental states. Self-ascription is done via a method that can in principle be used to form first-order beliefs. The method in this reading

[40] Depending on what exactly the special method is, it might still be more economic than inner-sense accounts.

is distinct from forming a first-order mental state, but it is not limited in its use to self-ascriptions and therefore not distinctive to self-knowledge. For example, suppose you form a first-order belief that there is a bottle on the table by perception. The strong reading of sameness has it that your self-ascription of that belief is also done by the same process of perception. The weaker reading states that your method of self-ascription cannot come apart from your belief-forming by perception, but is not identical to it. The weakest reading requires that your self-ascription is done by a method that can in principle form first-order beliefs but need not have any relation to perception, e.g. reasoning. Because Evans does not provide decisive evidence for either reading, I consider all versions as transparency accounts.[41]

This gives us a working definition:

> Transparency is a feature of the process of forming beliefs about one's own mental states. Such a belief-formation involves some sort of attending to an outward phenomenon which results in self-ascription (knowledge in a good case) of one's own mental state. The outward phenomenon has to be *exactly the same* as involved in forming the first-order mental state that is self-ascribed. Moreover, the procedure attending to it has to meet one of the following conditions for sameness of procedure:
>
> a) The procedure is exactly the same as in forming the first-order mental state that is self-ascribed; or
> b) the procedure is latched onto first-order mental state formation; or
> c) the procedure is in principle able to form first-order mental states.[42]

Evans did not propose it in this general form, but I take it that the idea was not meant to be limited to beliefs and perceptual experience, which are the cases he explicitly mentions. I believe that he rather used these two cases as examples for propositional attitudes and non-propositional states.[43] However, the views built on Evans's idea can be more limited. Some are only intended to explain knowledge of a small set of mental state types. Moreover, no view on the table claims that transparency is the only way we can get self-

[41] I will discuss in more detail one version of the strong reading (Boyle (2011)) and two version of the weakest reading (Moran (2001), Byrne (2005)) later.
[42] Note that the last formulation of being "in principle able to form first-order mental states" may prompt generality worries. Byrne (2005; 2011) for instance takes his account to explain self-knowledge by a process of reasoning involved in epistemic rules. He takes this to be economic, because we accept reasoning as a belief-forming process anyway. One might worry here that it is unclear how 'reasoning' should be individuated.
[43] Byrne (2011) agrees.

knowledge. With this general characterisation in place we can look at ways to develop the idea of transparent self-knowledge.

3.4 To Move or Not to Move

Antonia Peacocke (2017) recently proposed to distinguish between 'move' and 'no-move' accounts of transparency. The former are defined by a move from a judgment that p to a self-attribution of a belief that p that is epistemically warranted, whereas the latter deny any such move. Peacocke's description of this distinction might fit some accounts proposed recently (e.g. certain readings of Byrne (2005) and Moran (2001)), but it is not suitable for a general distinction. There are accounts that are structurally similar but do not fit the classification purely because they reject that the starting point for a transparency story of self-ascription is a judgment that p. Fernández (2013) for instance proposes an evidential state as the starting point for self-ascribing a belief and evidential states are quite different to judgments that p. Moreover, on a different reading of Moran the starting point is not a judgment, but merely the reasons that might be used in making a judgment.[44] Nevertheless, there is an important difference that Peacocke points to. Some accounts propose a distinct method that provides a move from the outward phenomena to a belief about one's belief, and some do not. This lines up with the different ways to understand the 'sameness of procedure' claim. The central difference is whether one thinks of first-order mental states and corresponding second-order beliefs as being generated by one, or by two processes. The strong and the weaker reading take it to be essentially one process, whereas the weakest reading proposes a distinct procedure of second-order belief formation.

One may object that the reading of the second-order process being latched onto the first-order process also accepts two processes. There is a sense in which this is true. However, the second-order belief-forming method in this case always entails a first-order belief formation. There is no method of self-ascribing a mental state that is (completely) distinct from a first-order process. Only in the weakest reading the self-ascription is done by a procedure that can be used independently for belief formation. It is used independently when one generates a first-order belief by said method.

We can now reformulate a general definition of move accounts:

[44] An overview of this reading of Moran will be given later.

> An account is a 'move' account iff it involves a move from an outward phenomenon to a self-attribution of a mental state by a *distinct procedure* (the transparency method) satisfying the weakest reading of 'same procedure.'

Any transparency account that denies this can be classified as a no-move account. This does not imply that no-move views do not involve any transition from the outward phenomena to a self-ascription of a mental state. Far from that, any transparency account has a transition of this kind. The crucial difference is the nature of the transition. Move views take the transition to be a distinct process from the generation of a first-order mental state. They can be understood as proceeding in two steps, so to speak. First forming the first-order state and then, second, forming the self-beliefs.[45] No-move views on the other hand take the transition to be sufficiently linked to the process of forming a first-order mental state. First-order states and self-beliefs are generated in a single step, so to speak. As I am going to show later this link can be spelled out in different ways. It can be located either at the formation process, or at the mental state.

With this broad distinction in place I can start analyzing the structure of transparency accounts. First I provide an overview and taxonomy of move accounts, showing how the analysis fits Moran (2001; 2003) as a case study. I will then provide a similar taxonomy for no-move accounts and use it to analyze Boyle (2009; 2011).

3.5 A Taxonomy for 'Move' Accounts

Move accounts emphasize that the transparency method is a distinct procedure that we can engage in. When one wants to find out what one believes, one attends to an outward phenomenon and in doing so one makes a move from this phenomenon to a self-ascription of a mental state. It is useful to describe these accounts based on their components:

1. What is the outward phenomenon that is the basis of the transition?
2. What is the move in question?
3. What is the result of the transition?

I will take these questions in reverse order. The third question seems to be the easiest to answer. Ideally the result is knowledge of a mental state. I understand knowledge here in line with the common notion of a justified true belief, plus a condition that rules out cases

[45] This does not imply that the second-order belief is based on the first-order state.

of relevant epistemic luck[46]. However, not all features of self-knowledge need to be explained by reference to transparency. Transparency might only by a partial explanation of self-knowledge in need of an additional explanation that does not rely on transparency at all. For instance, you may combine a transparency story of belief formation with a coherence account of justification. In this case the procedure explains why you have the belief, but it does not justify the belief.

With regard to the second question we can classify possible answers into at least three groups:

Inference moves take outward phenomena to be the basis of an inference towards a self-ascription, where inference need not be understood as deductive inference. Gallois (1996) for instance uses so-called Moore-inferences, a type of inference that is neither deductively valid, nor inductively strong. However, it is still warranted because the negation would be absurd, as it would generate a Moore-paradoxical thought or statement.[47] Byrne (2005; 2011; 2018) uses epistemic rules that are (practically) self-verifying in first-person formulation.

Deliberation moves take outward phenomena to be the basis of an agential process of weighting reasons for and against a proposition being true. The result of such a deliberation process then entitles one to a self-ascription. The paradigmatic example for this kind of move is Moran (2001)

Quietist moves claim that the outward phenomena are the basis for self-ascribing mental states, but they remain silent on how exactly this is done. An example of a quietist move is Fernández (2013).

Finally, – answering the first question – the outward phenomenon is usually construed as some sort of *outward directed mental state*, a *proposition*, or a *judgment.* The paradigmatic outward directed mental states are *justifying state*s.[48] Justifying states are outward directed insofar as they indicate the state of the world. If one attends to such a state one thereby attends to the world by proxy. One is not directly related to the world, but by mediation of a justifying state one can be related to the world. For instance, a visual

[46]Luck concerning the truth of the belief, i.e. *veritic* luck. For a discussion of epistemic luck see Pritchard (2005; 2007).
[47] For Moore's paradox see Moore (1951).
[48] E.g. Moran (2001), Fernández (2013)

experience can be a path towards a judgement about the world.[49] The visual experience itself does not include facts about the world itself, but it can be used to make claims about the world. The visual experience still gives you a positive epistemic standing towards a state of the world. It is a guide to truths – a guide to the world.[50] Justifying states are a possible outward phenomenon exactly because of this relation to the world. Moreover, it is unclear whether mental states other than justifiers can be outward directed in the right way and serve as an outward phenomenon. Fernández (2013) argues that at least urges can also be used as the starting point for a transparency story for self-knowledge.

An alternative proposal for an outward phenomenon focusses on propositions. One attends to the world in assessing whether a proposition is true or false. The assessed proposition is then used to form a second-order belief. One might favour propositions as an outward phenomenon if the justifiers alone cannot provide a basis suitable for the transition to a second-order belief. For instance, if you take your justifier to be a perceptual seeming then this might not be suitable for an inference to a self-ascription. However, the perceptual seeming can support the acceptance of a proposition which in turn is used as the basis for an inference. Using propositions as the outward phenomena has the further advantage that it seems to fit better into an externalist epistemology, because it does not require any mental, nor accessible justifiers that one attends to. Proponents of this idea include Gallois (1996) and Byrne (2005; 2011; 2018).

The final option found in the literature proposes that judgments are the outward phenomena. The idea here is that the act of judging that p provides a basis for a self-ascription, instead of the assessed proposition p. Judging is conceived as assessing the world and insofar as outward directed. Silins (2013) proposes that a judgment that p provides prima facie, propositional, introspective, immediate justification for the belief that one believes that p. Propositional justification here means that a subject has sufficient reason to believe something, even though the subject might actually not believe it (or believe it for different reasons). Silins does not provide a story how this propositional justification can be the basis for a belief, that is, how the propositional justification turns into doxastic justification. Because Silins' explanation is incomplete it is not all that clear to me whether judgment can satisfy the working definition of transparency that I proposed

[49] The paradigmatic example of using perceptual evidence as a starting point is Fernández (2013).
[50] Excluding skeptical concerns.

earlier. One of my requirements was that forming a second-order belief involves exactly the same outward phenomenon as forming the corresponding first-order belief. If the whole judgment is the outward phenomenon for the second-order belief-formation then this requirement cannot be met, because the whole judgment that p cannot be the outward phenomenon for forming the belief that p. After all, the best way to understand the judgment is that it is the belief-formation, and it cannot be the outward phenomenon for itself. Clearly Silins needs a different way to understand judgments – and he provides one. He talks of judgment as "a guide to belief" (Silins, 2013, p. 293) and of conscious judgments as mental actions which „[…]modify what it is like to be you at the time you perform them" (Silins, 2013, p. 294). His paradigm case of judgment is a case of sincere assertion that p (Silins, 2013, p. 294). I find this notion of judgment puzzling, and the paradigm case not particularly helpful. Moreover, if the relevant judgment is a conscious mental act, it also seems dubious whether it can be exactly the same outward phenomenon as is involved in forming first-order beliefs. The first-order beliefs do not appear to be formed in a way that involves conscious mental acts. What one might be conscious of are reasons that one attends to, but then these reasons as justifiers would be the outward phenomenon, and not the whole judgment. As is I suggest being sceptical of whether a judgment can be the outward phenomenon for transparency procedure.

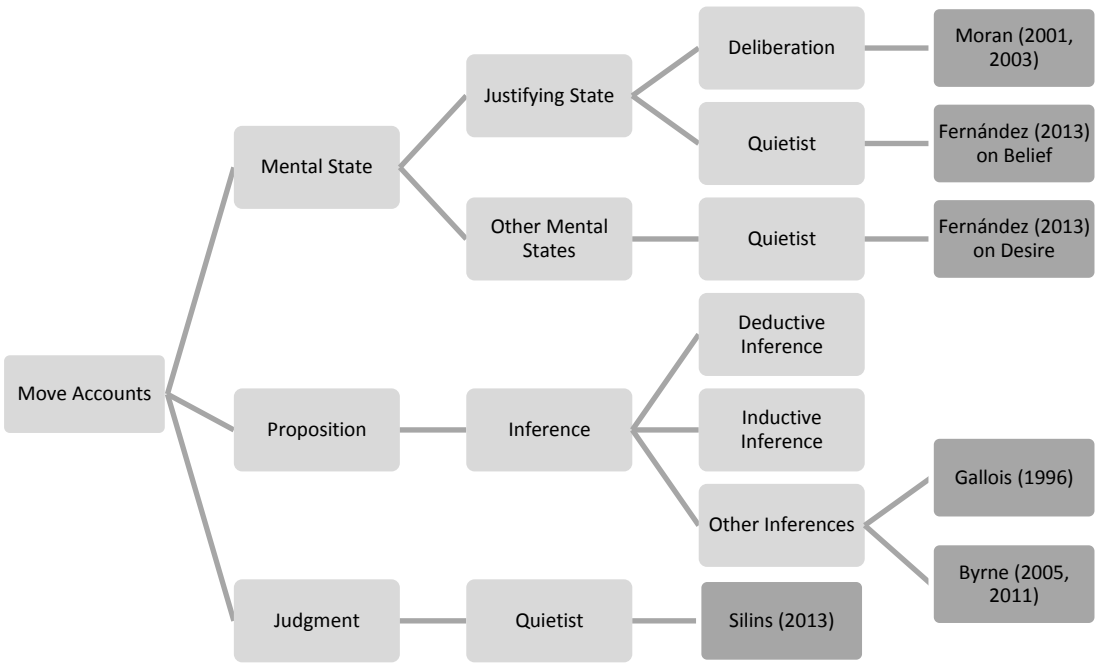

Figure 1 – A non-exhaustive overview of current move approaches

3.5.1 A Case Study - Richard Moran: Justifying State + Deliberation → Knowledge

Moran's transparency account is primarily built around agency for one's own mental realm. The aim is to provide an alternative to the mere theoretic, psychological approach that claims we observe our mental states while leaving them unaltered. This alternative involves self-knowledge best described as making up one's mind.[51] This is an action we can (but need not) undertake deliberately. When one asks "Do I believe that p?" one asks not with the aim of detecting an existing belief, but rather to start an inquiry of whether one ought to believe that p. The theoretical question whether one believes p is deferred to the deliberative question whether one ought to believe p. The latter question is then answered by attending to reasons for or against p. These reasons constitute the outward phenomena for our taxonomy of transparency. One way to understand the relation between the reasons for p and the self-ascription of a belief is that the question "Do I believe that p?" is deferred to the question "Am I to believe that p?" Because the latter is answered by attending to reasons that justify an answer, the former is also answered by attending to such reasons. Hence, the outward phenomena are reasons as justifying states.[52]

In Moran's framework we can attend to the outward phenomena by mobilizing and engaging with reasons in a deliberative stance. Furthermore, self-reflection under a deliberative stance ends in an act of agency - in a decision or a commitment of some sort (Moran, 2001, p. 58). By engaging with reasons in a deliberative stance one can self-ascribe a belief that p and at the same time make up one's mind about believing p. One commits to believing p, if one's reasons are in favor of p being true. Moreover, one can self-ascribe the belief that p under the same condition. This is Moran's way of spelling out the idea of attending to the outward phenomena.

Finally, what about the resulting state in Moran's picture? One starts by asking "Do I believe that p?" then defers to the question "Am I to believe that p?" In the next step one deliberates on this question based on all available reasons. Suppose they are in favor of p being true. Then one ends the deliberation by both believing that p, and self-ascribing that

[51] He allows other ways of theoretical self-knowledge, e.g. in case of therapeutic self-interpretation (Moran, 2001, p. 85)

[52] See Moran (2001, p. 63). This way of understanding Moran's transparency account is also proposed by Finkelstein (2012).

one believes that p. The process makes it the case that one believes p, and it is also the basis for the self-ascription. But is this self-ascription justified? Prima facie it does not seem to be justified. The deliberation engaged with reasons for and against p, not reasons for or against having the belief that p.

However, in Moran's conceptual landscape there is a solution available, which he develops in a response to O'Brien and Shoemaker (Moran (2003)). There is justification for self-ascribing by deliberating on reasons for p. The thought behind this is that one has to take oneself to be a rational agent that bases one's beliefs on reasons. If one does not, then the whole idea of justification collapses because the gap between reasons and beliefs could not be closed at all (Moran, 2003, p. 406). Given that one can justify beliefs, one is a rational agent, and therefore beliefs are based on reasons. Hence, one can justifiably self-ascribe a belief that p based on reasons for p. The transition from reasons for p to self-ascribing the belief that p comes with being a rational agent. And we are entitled to assume we are rational agents because otherwise justification would be impossible in general.[53]

With this argument we have a resulting state of justified belief, but is there also a case for truth of the self-ascription? If one actually makes up one's mind based on the reasons for p, and one self-ascribes based on the reasons for p, and, finally, one is a rational agent, then the self-ascription turns out to be true. However, one might fail on any of these steps in a particular instance. One might be the victim (and perpetrator) of self-deception and self-ascribe based on different reasons than those that actually determine the belief. Transparency can also fail if I cannot become aware of my belief by reflection because "[…] I simply won't reflect explicitly and steadily on its object" (Moran, 2003, p. 408). Self-reflecting based on transparency as Moran describes it therefore does not guarantee knowledge. However, in a good case it will result in knowledge. One arrives at self-knowledge (resulting state) by deliberating on (attending to/transitioning from) reasons for p (the outward phenomena). In a good case it is self-knowledge, because it is a justified true self-ascription. It is true, because making up one's mind ends with one having the ascribed belief. Moreover, it is justified in virtue of the transcendental argument based on what it means to be a rational being. Again, it does not guarantee self-knowledge, but every single part of the knowledge condition being satisfied in a good case of deliberation is explained in virtue of Moran's account.

[53] Gertler (2011a, p. 189) provides a fantastic reconstruction of this argument.

3.5.2 A Case Study – Alex Byrne: Proposition + Epistemic Rules → Justified Second-Order Beliefs

Byrne's (2005; 2011; 2012a; 2012b; 2018) account differs from Moran significantly. Whereas Moran highlights the importance of the agent and the deliberative stance, Byrne's proposal is more in line with a peculiar, theoretical method of coming to know one's mental states. On his view the transparency method is spelled out as an epistemic rule, that is, a rule of belief-formation. This kind of rule is best illustrated with Byrne's simple example of Mrs. Hudson's doorbell ringing:

> Mrs. Hudson might hear the doorbell ring, and conclude that there is someone at the door. By hearing that the doorbell is ringing, Mrs. Hudson knows that the doorbell is ringing; by reasoning, she knows that there is someone at the door (Byrne, 2005, p. 94).

This is translated as the rule "If the doorbell rings, believe that there is someone at the door." Here Byrne straightforwardly tells us, that Mrs. Hudson follows the conditional of the rule by reasoning, which is a distinct process from the perception of the doorbell ringing. Mrs. Hudson first recognizes that the doorbell is ringing. She then transitions by reasoning from the doorbell ringing to the belief that someone is at the door. Byrne proposes that a similar story can be given for forming beliefs about one's mental states. His paradigmatic example of a similar rule for belief is the following:

(BEL) If p, believe that you believe that p (Byrne, 2005, p. 95).

Just as in Mrs. Hudson's case, one can follow (BEL) by recognizing that p, and then making a transition to the belief that one believes that p. This transition is a process of reasoning. Moreover, the resulting belief is true because of one's recognition of p. The mere fact that I recognize that p makes the second-order belief true, because recognizing that p includes knowing that p, and hence believing that p. Byrne labels this feature 'self-verifying' (2005, p. 95). This self-verifying nature makes (BEL) a *good rule* to follow. It is truth-conducive. Whenever I form a second-order belief by following (BEL) I am guaranteed to end up with a true belief. (BEL) is not merely reliable. It is, so to speak, hyper-reliable. Reliability would merely require a preponderance of true over false beliefs generated by a process, but the epistemic rule (BEL) produces *only* true second-order beliefs. Moreover, even merely trying to follow (BEL) generates true second-order beliefs. Because (BEL) provides even more than required for reliability this also entails that the second-order beliefs are formed reliably and hence are prima facie justified.

(BEL) is a transparent rule, because the starting point for the rule is an outward phenomenon – the proposition p – not a mental state. This outward phenomenon is the basis for the transition to a self-ascription of a belief. Moreover, the transition is understood as a form of reasoning, and therefore fits with the weakest reading of 'sameness of procedure.' It is a procedure that in principle could be used to form first-order beliefs. For this reason Byrne can rightly point to his view as especially economic: His account requires no belief-formation process that is not accepted independently anyway. For instance, one recognizes that a particular car is red by perception, and then comes to believe that one believes that the car is red by reasoning according to the epistemic rule (BEL). All the belief-formation processes required here are perception and reasoning, both of which we are happy to accept independently of (BEL).

For a complete account of self-knowledge a single rule for forming beliefs about one's beliefs will not be enough. Therefore, Byrne aims to provide analogues to (BEL) for other mental states, such as intention (2011), perception (2012a), desire (2012b), sensations such as pain (2018) and even memory (2018). I take this to be a response to shortcomings of transparency accounts such as Moran's (2001) in explaining how one comes to know one's mental states other than belief. This issue will be revisited in part 7.

3.6 A Taxonomy for 'No-Move' Accounts

After discussing move accounts we can now look at the alternative. No-move accounts are views that include an essential link between the formation of a first-order mental state and a self-ascription of that state. They satisfy the notion of transparency because the link ensures that self-ascriptions are done by attending to outward phenomena. The link can ensure this insofar as a first-order mental state is generated by attending outwards and the link transfers this outward direction over to the self-ascription. The term 'link' here ought to be understood as a metaphysical claim, not a mere epistemic metaphor. With this proposal no-move accounts need not say much about the outward phenomena in question, because whatever the phenomenon with relation to forming first-order mental states is, it performs the same role for self-ascribing these states.

We can characterise no-move accounts based on their answer to three questions:

1. Where is the link between the formation of a first-order mental state and self-ascription of that state located?
2. What is the nature of this link?

3. What does the link explain?

The answer to (3) is similar to the one given for move accounts: ideally it explains knowledge of one's mental state, but it need not. It is in principle possible that the link explains the formation of a belief about one's mental state without explaining itself why that belief is justified or true.

I will discuss (1) and (2) together, because the answer to the first question sets the stage for the second one. With regard to (1) the link in question can be located at two different levels: by a link on the level of the process of forming mental states, or on the level of the formed mental states. These are not exclusive, so it is possible to have a link on both levels. Moreover, any link on the level of formed mental states entails a link on the process level, but not the other way round.

Linked processes are such that the formation of first-order mental states either entails or generally comes with the formation of second-order beliefs about these first-order states (or the other way round). Accounts using the idea of linked processes involve metaphysical claims about the nature of certain processes generating mental states, which provides an answer to (2). An example of a transparency account in virtue of linked processes is provided by Peacocke (2017).

Linked states are such that the first-order mental state is identical to or entails[54] the second-order belief about that state (or the other way round). Accounts using the idea of linked states involve metaphysical claims about the nature of certain mental states, which provides an answer to (2). A link in mental states entails a link in generating processes. If a first-order mental state and the second-order belief about that state are identical then both are generated by the same process. Moreover, if a first-order mental state entails a second-order belief then the first-order process puts one in a position to engage in the second-order process. Either way, the link in states comes with a link in processes. The paradigmatic instance of a linked states account is Boyle (2009; 2011), who proposes that beliefs are always tacitly known insofar as a belief and knowledge of this belief are only a single mental state.

[54] Note that 'entailment' here is taken to be the involvement of a necessary consequence. A first-order mental state entails a corresponding second-order belief if this second-order belief is a necessary consequence of the first-order state.

Both processes and states can be linked at two stages, either at the first-order stage or at the second-order stage. The link of processes occurs at the first-order stage in case the first-order process entails or generally comes with formation of second-order beliefs.[55] It occurs at the second-order stage if the second-order process entails or generally comes with a first-order process.

Similarly, the mental state link is at the first-order stage if the first-order mental state is identical to or entails the second-order belief, and it occurs at the second-order stage if the second-order belief entails the first-order mental state. It cannot be identical at the second-order stage because identity is a symmetrical relation, whereas entailment is not.

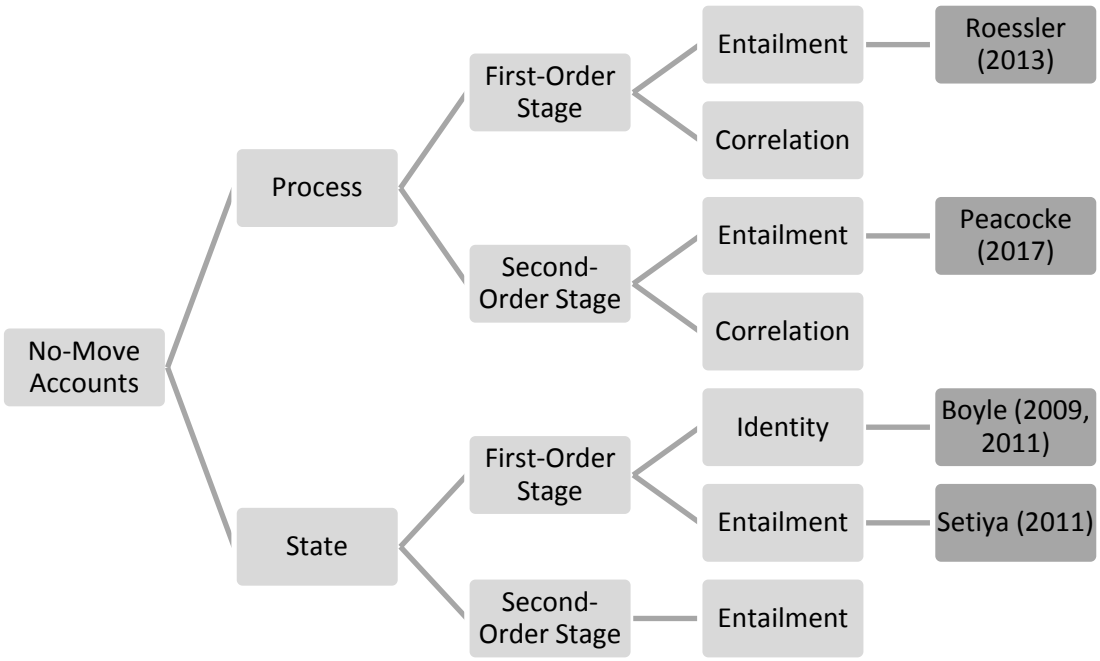

Figure 2 – An overview of the no-move landscape

3.6.1 A Case Study – Matthew Boyle: Linked Mental States at the First-Order Stage

Boyle (2009; 2011) aims to explain transparent self-knowledge in a way that is supposed to capture essential features of Moran's account without falling into the problems of a distinct

[55] For instance, Roessler (2013) proposes that in intentionally judging that p (a process at the first-order stage) one also gets an awareness of one's belief that p

move from an outward phenomenon to a self-ascription of a belief. Boyle takes it that such a move is not justified, but furthermore, it is not needed. We can avoid this requirement. Hence, we do not need a transcendental argument as Moran (2003) discusses, or a reasoning-based move as Byrne (2011) employs.[56] Instead, Boyle proposes a hypothesis on the nature of belief that is supposed to explain why self-knowledge functions based on transparency. The core hypothesis is that the following is a basic, irreducible fact about human believing:

> [A] subject in this condition [believing] is such as to be tacitly cognizant of being in this condition. Hence in the normal and basic case, believing *p* and knowing oneself to believe *p* are not two cognitive states; they are two aspects of *one* cognitive state—the state, as we might put it, of knowingly believing *P* (Boyle, 2011, p. 228).

The idea is that a belief is by nature also tacit knowledge of the belief. This tacit nature allows one to be unaware of one's own belief. Boyle does not doubt that we are not always aware of our beliefs. However, he can explain this without splitting up first- and second-order beliefs by using this notion of the second-order belief being tacit. The second-order belief is a different aspect of the first-order belief, one that we can become aware of by reflection. Importantly, reflection is not a belief-forming process. It is only a process of becoming consciously aware of a tacit state. Whenever one forms any belief one also forms tacit knowledge of this belief at the same time. Thereby Boyle satisfies transparency insofar as "I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (Evans, 1982, p. 225). This is trivially satisfied because the same procedure that produces the belief also produces knowledge of that belief. They are the same state, so if a procedure produces one, it also produces the other. Furthermore, the state is generated by attending to an outward phenomenon because the first-order belief is formed by attending outwards. For instance, when I see a magpie sitting on a fence I form a belief by perception, which is obviously outward directed. Given that this very belief is also knowledge of the belief this knowledge is generated at the same time in the same way. Hence, knowledge of the belief that there is a magpie sitting on a fence is generated by attending outwards.

Given that Boyle takes first-order belief and knowledge of that belief to be a single state, he also denies a distinct second-order belief-forming process. This is an instance of linked

[56] A dedicated criticism of Moran and Byrne based on this line of argument is provided by Antonia Peacocke (2017).

states entailing linked processes. Hence, when you form a first-order belief, you also get tacit knowledge of this belief for free, according to Boyle. The first-order belief formation does both jobs: generating a belief and the tacit knowledge of that belief. Moreover, because the states are identical the explanation does not require a distinct move from an outward phenomenon to a self-ascription.

3.7 Using the Taxonomies as Diagnostic Tools

We can now use these taxonomies to understand why some kinds of transparency accounts are confronted with specific objections and criticism. Moreover, we can find out how to avoid these objections. I focus on two problems as follows:

1. How can transparency account for self-knowledge of attitudes other than belief? (The problem of scope[57])
2. How can transparency account for self-knowledge of stored beliefs? (The standing state problem)

3.7.1 The Problem of Scope

A common charge against transparency accounts of self-knowledge is that they are limited to belief. This criticism has been leveled particularly against Moran's version of a transparency account (cf. Finkelstein (2003), Gertler (2011a), Cassam (2014)). There are some sociological reasons for this. However, there are also reasons intrinsic to the account that point to the problem of scope.

First, Moran explicitly endorses some uniformity for explaining self-knowledge. He distinguishes two categories of psychological states we can self-ascribe in a privileged way: occurrent states such as sensations and passing thoughts and standing attitudes, such as belief, desire, emotional attitudes, and intention. His aim is to say something about knowledge of all types of standing attitudes, but not of sensations (Moran, 2001, pp. 9-10).

Second, Moran endorses a framework of transparency that makes it difficult to see how his account can explain knowledge of all types of standing attitudes. It is easy to find individual counterexamples. Take Finkelstein's (2003, p. 163) example of wanting to know whether I adore my dog.[58] Moran's picture suggests that we can relate the question of whether I adore my dog to reasons that tell me that I should (or should not) adore my dog. However, intuitively there are no rational obligations to adore my dog whatsoever. A related line of

[57] The notion is taken from Gallois (1996).
[58] A similar example involving an irrational fear of a spider is developed in Finkelstein (2012).

argument against transparent self-knowledge of emotions has been brought up by Kloosterboer (2015).

Individual counterexamples are useful to put pressure on Moran's account, but they do not teach us much of a general lesson. However, I take it that we can use the taxonomical analysis of Moran's account to locate the source of the problems with his position. The difficulty for Moran lies in his narrow notion of outward phenomena. His notion of transparency starts with phenomena that are distinctly epistemic. They start with epistemic reasons.[59] These are starting points that are directly and essentially related to justifying first-order beliefs. More generally, epistemic reasons are justifiers, and therefore well suited to explain self-knowledge of belief. However, there is no obvious connection between a justifier and mental states other than belief – mental states that are not in the justification game. Why should one's attending to a justifier bring one any closer to self-ascription of states that do not care about justifiers? Why should a justifier be connected to emotions? There seem to be no good reasons for any such connection – or at least Moran does not give us any. There might be a way to allow the inclusion of practical reasons that connect to desires and actions[60], but that will not be enough to find an explanation for self-knowledge of all types of mental states. For instance, practical reasons are also not sufficiently connected to emotions or phenomenal states.

However, this is not a problem limited to Moran. Whenever a transparency account makes claims about the nature of the outward phenomena these phenomena need to be related to the mental states which can be known by transparency. Having a single account that explains knowledge of different kinds of mental states requires outward phenomena that are related to all these different kinds of states. This is a problem for views that use justifying states as outward phenomena like Moran does.

[59] In Moran (2003) he makes this explicit: "The more general and central truth in this context is that I answer the external question about the weather or the possibility of war by putting myself in a position to confront and assess the reasons relevant to the truth about the weather or the possibility of war. In some cases, these reasons will be so immediate and obvious to me that I don't even have to think of them as reasons. (It may never have occurred to me to think otherwise, I've always assumed it and everything else speaks in favor of it and nothing against it, etc.) So I think the Transparency claim is better put by saying: When asked 'Do you believe P?', I can answer this question by consideration of the reasons in favor of P itself" (Moran, 2003, p. 405).

[60] Moran endorses this option when discussing responsibility for desires that are the conclusion of practical reasoning (Moran, 2001, p. 117).

Proponents of transparency have at least four different responses to the problem available. First, a divide and conquer strategy that accepts that knowledge of different mental states is based on different outward phenomena. Fernández (2013) tries this path with a different outward phenomenon for desires. He argues for symmetry in structure between evidential states causing (and justifying) beliefs and mental states causing desires (and justifying beliefs about them). Just like evidential states are outward phenomena for self-knowledge of beliefs, mental states causing desires can be the outward phenomena for self-knowledge of desires. I have some doubts whether he succeeds. He proposes a 'production-of-desire' principle that links the production of desires to either (i) other desires, (ii) experienced urges, or (iii) a subject valuing that p. (Fernández, 2013, p. 84) However, it is unclear what 'urges' exactly are, why they cause desires, and why they cannot be caused by desires. (Ashwell, 2013b) Moreover, it is also unclear how we should understand valuing without reference to desiring. The worry here is that (iii) becomes derivative to (i), which leaves the only independent cause of desires to be the problematic urges.

Second, one can opt for an account with outward phenomena that are connected to various types of mental states. Propositions seem to be a candidate as their content can be diverse enough to be related to self-ascribing different mental state types. Alex Byrne proposes this kind of account in a piecemeal fashion, trying to show that it fits for belief (2005; 2011), intention (2011), perception (2012a), desire (2012b) and sensations (2018). His opposition claims that he fails to do so (Ashwell, 2013a; Samoilova, 2016), but Byrne (2018) provides responses to some of their objections. There is at least no obvious reason why Byrne's approach cannot succeed. In contrast, starting with justifying states seems a liability from the very start and unable to solve the problem of scope.

Third, one can argue for a more limited scope of explanation. One can concede that transparency cannot explain attitudes other than belief, but emphasize that it is the best explanation for knowledge of one's beliefs. Boyle (2009; 2011) argues for this from a no-move account's point of view. However, there is nothing in principle that limits this response to no-move accounts.

Fourth, one can opt for quietism regarding the outward phenomena and thereby evade the problem of scope to some extent. This option is only available for no-move accounts which do not require a transition based on an explicit outward phenomenon. However, this is not to say that all no-move views can solve the problem of scope. To the contrary, a majority

does not avoid the problem at all insofar as these accounts need to show that the link between first- and second-order processes or states is present for various types of processes or states. This is an especially high hurdle for linked states accounts which need to show how different mental states either entail or are identical to second-order beliefs about these states. It is comparatively easier for linked process views[61] to solve the problem of scope. The least demanding option is an account of linked processes that requires only a general correlation of processes, but no entailment. Such an account proposes that forming first-order mental states generally goes together with the formation of a corresponding second-order belief, but denies any entailment relation between first-order and second-order state. This account denies conceptual relations between the two processes, but proposes a link as an empirical hypothesis instead. Hence, the link can be present for all (or many) mental state types. Any such account avoids the problem of scope by placing the burden on the empirical sciences.

3.7.2 The Standing State Problem

The second objection is the standing state problem, put forward by Shah and Velleman (2005), Gertler (2011a) and Cassam (2014). The concern is that transparency cannot explain how one knows one's stored states. Self-ascribing a mental state by transparency involves attending to outward phenomena that are the basis for these mental states. Hence it seems at least plausible that attending itself prompts the (re-)formation of mental states. If that is true, then there is no way to make sure that I self-ascribe a belief that I already had before I started to form a second-order belief by transparency.

Consider the following example: Aimee believes that she is 31 years old. She wants to know whether she believes that she is 31 years old. How is she supposed to acquire this knowledge? If she considers her reasons for being 31 she will end up with a belief, but she has no way to tell whether this is a belief she already had before or a new one. The only way to be sure Aimee finds out what she already believes is if she were somehow able to deliberate on her age without this deliberation influencing her state of believing. In addition, the reasons available need to be identical with those which were available when she initially formed her belief.

[61] However, these depend on the process in question. Some linked process view fail to solve the problem of scope because their starting points are only intentional processes (intentional actions). Roessler (2013) seems to fall into this category.

To solve the standing state problem we need to locate its source, and the taxonomy of transparency accounts can help us out here. First, notice that the standing state problem occurs in case we posit two distinct processes of mental state formation. It requires that we can have a mental state M at t_1 and then at t_2 we aim to self-ascribe M. In doing so we are at risk of changing M to M*, because self-ascribing a mental state involves attending to outward phenomena which can prompt the generation or change of a mental state. This picture requires two distinct processes, one at t_1 producing the mental state, and one at t_2 self-ascribing and at the same time possibly (re-)forming a mental state. This requirement is met by all move accounts. Remember that move accounts are defined as having a distinct procedure (the transparency method) satisfying the weakest reading of 'same procedure.' In virtue of this the transparency method is always separated from the initial formation of the first-order mental states at t_1. Moreover, because they still require outward phenomena as a starting point they inherit the risk of letting these phenomena prompt a (re-)formation of mental states at t_2. The precise nature of the outward phenomena makes no difference here, as all move account accept this general structure.

There are at least two possible responses here. First, limit the explanatory aim of one's transparency account. Gertler (2011a) recommends this for rationalist transparency proponents, e.g. Moran (2001). They should deny that transparency ought to explain all kinds of self-knowledge, but argue that transparency explains critical self-knowledge. Critical self-knowledge is here understood as knowledge which's justifying reasons have been evaluated (Gertler, 2011a, p. 169). Not all attitudes we have are based on such reasons. However, as reasoners in general we ought to aim for the ideal of being a critical reasoner. We ought to have reasons for our attitudes, and self-knowledge gives us the possibility to make sure we reach this aim. Moreover, self-knowledge as a way to critically engage with our own attitudes is a precondition to being open to certain forms of criticism in case we don't base our attitudes on reasons. If one accepts this rationalist picture, then it is a plausible move to restrict the relevance of transparency to attitudes one currently holds. Hence, Gertler suggests that Moran should be fine limiting his explanatory aim to one's current attitudes (Gertler, 2011a, p. 19). However, if one wants to stick to a more general explanatory aim this solution is insufficient.

The second response is to opt for a no-move account. However, not every no-move account will do the trick. The same gap between forming a mental state at t_1 and possible (re-)

forming the mental state at t_2 occurs if we take the link to be at a second-order stage. Both linked processes and linked states at a second-order stage suppose that you can form first-order mental states without a second-order belief about these states, but they claim that you cannot (or generally do not) form a second-order belief about a mental state without forming that mental state. Hence second-order stage no-move accounts still involve the possibility of forming a mental state M at t_1 and then when self-ascribing the mental state at t_2 changing M to M*. Only first-order stage links can avoid the standing state problem. They can do so because they deny that forming first-order mental states is generally distinct from forming second-order beliefs about these mental states. For instance, remember Boyle's (2009; 2011) proposal that a belief is at the same time knowledge of that belief. Given this assumption it is impossible to have a belief that p without having implicit knowledge of that belief. There is no room for the belief without knowledge of the belief, and therefore no room for a standing belief that is unknown. This solution is not limited to linked state proposals. For if you were to suppose that every first-order state production comes with the corresponding second-order belief we could also know standing states just fine. There would not be a mental state M *before* introspection, because forming M and introspecting would not be distinct processes.

3.7.3 Mapping out the viable options

Based on my discussion of the problem of scope and the standing state problem we see how we can use my proposed taxonomies to find better options for transparency accounts. Once we map out the landscape of possible accounts we form an explanatory goal. We may aim for explanations on a scale between a broad, unified explanation of all mental state types, or, on the other end, an explanation restricted to critical self-knowledge of belief. Once we set our goal we can eliminate specific views and complete paths of views that are unable to reach this goal. I aim for an account that deals with the standing state problem, so I can eliminate all move accounts and all second-order stage no-move accounts. Moreover, I want the view to also deal with the problem of scope so I can further eliminate no-move accounts that employ linked states. I quickly end up with a narrow scope of live possibilities to build a new, viable transparency account: *A no-move account based on linked processes on the first-order stage*.

3.8 Conclusion

In this chapter I clarified what it means for any account of self-knowledge to be a transparency account. I started with the general observation of Evans (1982) and identified the important criteria for transparency views of self-ascription: The formation of first-order mental states and the self-ascription of these states need to be done by attending to the same outward phenomena and by using the same procedure. I clarified how to understand 'same procedure' and provided a general distinction between move and no-move accounts. Moreover, I presented a taxonomy that captures the landscape of transparency accounts. I have no doubts that it is incomplete, but I am confident that it captures the most prominent views in current literature. Finally, I used these taxonomies to get a better grasp on common problems for transparency accounts. We can understand the problem of scope and the standing state problem better with an overarching background of the structure of possible transparency accounts. Moreover, we can find new and better options by understanding this structure. I suggest that if we want a unified transparency account that avoids the standing state problem we should look into no-move accounts that link processes at a first-order stage. This is plan for the next chapter.

4 The Single Process Model of Self-Knowledge

In this chapter I provide a model of self-knowledge that accounts for mental state self-ascriptions in virtue of a single process generating both a mental state and a second-order belief about this mental state. The account explains the apparent difference between getting to know our own mental states and the mental states of other human beings. Moreover, the model provides a plausible story for the reliability and fallibility of self-ascriptions. This is accomplished in virtue of novel take on a transparency account inspired by Evans (1982).

I start by setting up the explanandum and then propose a no-move account via linked processes at the first-order stage. I then provide a core principle of the single process model, claiming that mental state self-ascriptions and the corresponding first-order attitudes are produced by a single process. I spell out important implications of this principle and further refine it to solve immediate problems. Furthermore, I show how the principle gives us a convincing story to tell for all features set up as the explanandum. The single process model has an explanation for the apparent asymmetry, reliability, fallibility, and also the transparency observation. Finally, I consider some problems for the account.

4.1 Introduction

In this chapter I argue for a model of self-knowledge that accounts for mental state self-ascriptions in virtue of a single process generating a mental state and a second-order belief about this mental state. The account explains the apparent difference between getting to know our own mental states and the mental states of other human beings. Moreover, the model provides a plausible story for the reliability and fallibility of self-ascriptions. This is accomplished in virtue of a different take on transparency inspired by Evans (1982). My goal here is relatively modest: I aim to show that the single process model of self-knowledge is explanatory powerful and economic, and hence should be taken seriously. Whether it actually is correct will ultimately be decided by the sciences, but I provide reasons why the model is a good candidate to explain self-knowledge.

I start by setting up the explanandum and propose a no-move account via linked processes at the first-order stage. I provide the core principle of the single process model, claiming that mental state self-ascriptions and the corresponding first-order attitudes are produced by a single process. I spell out important implications of this principle and further refine it to solve immediate problems. Furthermore, I show how the principle gives us a convincing story to tell for all features set up as the explanandum. The single process model has an explanation for the apparent asymmetry, reliability, fallibility, and also the transparency

observation. In the final section I consider some potential problems for my view and conclude with an outlook on what still needs to be done to further develop the account.

4.2 Setting the Stage

Any account of self-knowledge has to explain a set of features that we ascribe to the phenomenon in question. In chapter 1 I discussed how our identification of these features is based on observing our linguistic practices. Moreover, in chapter 2 I argued that these features should be spelled out on the level of belief according to the doxastic view. These features are:

- ***Asymmetry***: Beliefs about one's own mental states are usually[62] formed differently than beliefs about other's mental states or the external world.
- ***Reliability*****:** Beliefs about one's own mental states formed in this peculiar way are usually more reliable than beliefs about the external world or other people's mental states.
- ***Fallibility*****:** Beliefs about one's mental states can be wrong.

Asymmetry and reliability explain our ordinary linguistic practices. Normally, avowals are reports of *peculiarly formed* and *highly reliable* beliefs, which explain why usually avowals are interpreted as being true. There might be overriding reasons to challenge or question an avowal (e.g. when one has especially good evidence that an avower is self-deceived; when one has reasons that the avower is insincere, or lacks the right linguistic concepts), but in ordinary circumstances an avowal of an adult has the presumption of truth. Asymmetry and reliability explain this presumption.

Reliability should not be read as infallibility. I accept that one can be wrong about one's own mental state, as I already assumed when discussing mental state disagreements in chapter 2. Failure of self-knowledge is possible. This is supported by empirical data such as studies on the unreliability of verbal reports shown by Nisbett and Wilson (1977), the experimental induction of beliefs that one intended to act a certain way by Wegner and Wheatley (1999), and the studies of split-brain patients analyzed by Gazzaniga (1995).

[62] In some cases one's own mental state might only be accessible via self-directed mind-reading. In these cases one observes one's own behavior and then infers one's own mental state.

Moreover, it is supported by thought experiments such as the classic case presented by Peacocke (1998):

> Someone may judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (Peacocke, 1998, p. 90).

Both the empirical studies and the thought experiment are challenged. Wilson (2002) himself argues against the significance of the empirical studies, Parent (2016) provides further rebuttals to empirical cases. Moreover, Parent (2007) and Burge (1988) defend restricted infallibilism based on compositionality principles of thoughts. I side with the fallibilists in full awareness that I am not providing any definitive argument to end this discussion here.

Furthermore, I take it that any account of self-knowledge has to fit with our phenomenology of the formation of self-beliefs. I hinted in chapter 3 that the phenomenology is not unified, because most cases are experienced as self-knowledge being formed silently – without any awareness of the process of forming second-order beliefs, nor awareness of the basis of such beliefs – whereas in other cases the formation of second-order beliefs is experienced as transparent to the process of forming first-order beliefs. Explaining the asymmetry ought to account for this difference as well.

Adding to these general features to be explained, I aim to provide a transparency account of self-knowledge. My motivation for this has been laid out in chapter 3 and consists in a commitment to an alternative to inner-sense that is more economic without losing out on explanatory power. A further desideratum of the account is to make sure this promise of a transparency account is fulfilled. Hence, the fourth feature to account for is transparency.

- ***Transparency*****:** Beliefs about one's own mental states are formed in a way that involves some sort of attending to an outward phenomenon which results in self-ascription (knowledge in a good case) of one's own mental state. The outward phenomenon has to be *exactly the same* as involved in forming the first-order mental state that is self-ascribed. Moreover, the procedure attending to it has to meet one of the following conditions for sameness of procedure:

a) The procedure is exactly the same as in forming the first-order mental state that is self-ascribed; or
b) the procedure is latched onto first-order mental state formation; or
c) the procedure is in principle able to form first-order mental states.[63]

Transparency as a feature is different from asymmetry, reliability, and fallibility. Transparency is not a feature that self-knowledge as the explanandum has, rather it is a constraint on what my proposal for an explanans should look like. I take the constraint on board with the aim of ending up with an especially good explanation for asymmetry, reliability, and fallibility. There is no guarantee that the constraint leads to the best possible explanation, but I show that it leads to one worthy of consideration.

Finally, luminosity, the idea that one is in a position to know that one is in a mental state simply by being in that mental state, does not show up in the features I aim to explain. In this chapter I focus on propositional and non-propositional attitudes taken as non-experiential. The evidence in favour of fallibility also indicates that these do not seem to be luminous. It seems perfectly possible to have an attitude without being aware of having the attitude. The biased administrator Peacocke's example above judges that she is fair to applicants, but her actions indicate that she believes that her compatriots are superior. Here she appears to have a belief, without being in a position to know that she has the belief.

Observing our linguistic practice also shows that we do not expect attitudes to be luminous.[64] For instance, in an early stage of a potential relationship one might be unable to tell whether one is in love with another person and it seems appropriate to answer 'I'm not sure' when asked. Similarly, it seems felicitous – though quite unhelpful – to answer 'I do not know whether I want ice cream right now' when a friend asks you before they are going to the ice cream parlour. Nevertheless, usually we can answer that we are in a certain attitude when asked. My proposed model is going to allow that. I will return to luminosity in chapter 5 when I discuss the single process model for nonintentional, experiential states.

[63] For more on this formulation of transparency see chapter 3.
[64] Wright (1998, pp. 16-17) agrees to some extent. He claims that we do not expect others to be able to avow their attitudes all the time, but we do expect it for some attitudes that he calls 'basic.' I believe that even for his basic attitudes there are cases in which the linguistic practice allows for claiming ignorance of having the attitude. The ice cream example at the end of the paragraph is a prime candidate for one of Wright's basic attitudes for which we can appropriately claim ignorance.

4.3 The Single Process Model: A No-Move Account via Linked Processes at the First-Order Stage

The transparency account I aim for should be able to deal with both the problem of scope, and the standing state problem as discussed in chapter 3. Hence, I aim for a unified transparency account of self-knowledge. I provide this in two steps: an account for attitudes in this chapter; and an account for non-intentional mental states in the next chapter. This is mostly for ease of exposition, and not because there are fundamental differences in forming beliefs about one's mental states based on the type of mental state involved. The same structural principles will be at work in both cases, such that the promise of a unified account can be fulfilled.

Looking back at the taxonomy of available transparency accounts the aim of a unified account that also avoids the standing state problem limits the available options. Given that I take the standing state problem seriously, I ought to avoid move accounts, because they cannot avoid the problem. The standing state problem requires that we can have a mental state M at t_1 independently from any belief about that mental state M. Then at t_2 we aim to self-ascribe M. In doing so we are at risk of changing M to M*, because self-ascribing a mental state involves attending to outward phenomena which can prompt the generation or change of a mental state. This picture requires two distinct processes, one at t_1 producing the mental state, and one at t_2 self-ascribing and at the same time possibly (re-) forming a mental state. This requirement is met by all move accounts. If we want to avoid the standing state problem, we therefore have to avoid all move account options. My preferred way to avoid the problem is to deny that there are two independent processes going on: one that generates a mental state M at t_1, and another one aiming to self-ascribe the mental state M at t_2. A suitable no-move account establishes a connection between the generation of M and the self-ascription of M already at t_1, hence it avoids the standing state problem. However, any such connection is only available for first-order stage no-move accounts. Second-order stage versions fall into the same problem as move-accounts.

Furthermore, the aim of a unified transparency accounts also requires me to deal with the problem of scope. Hence, I can further eliminate no-move accounts that employ linked states. The reason for this is that it is highly unlikely that every mental state type is linked to a second-order belief about that mental state. While it might be plausible to link first-order

beliefs with second-order beliefs[65], there is no good reason to posit any such link between first-order desires and second-order beliefs, or first-order hopes and second-order beliefs. Moreover, even if there were such reasons, the cost of any such view would be an ontology of mental states that is rather counterintuitive, and has troubles explaining mental states of which one is unaware.

I take the best bet for an original, viable transparency account to be the following: *A no-move account based on linked processes on the first-order stage*. This path provides the most economic, unified transparency account and avoids the standing state problem. Moreover, I take the link to be one of correlation, not entailment. This is a more achievable goal than the stronger thesis that first-order mental state formation entails the formation of justified second-order beliefs. The only view that I am aware of which proposes this stronger idea is Johannes Roessler's (2013) proposal that any intentional act of judging that p itself justifies the self-ascription of the belief that p.[66] The main idea here is that an intentional act can only be intentional if it involves knowing what you are doing by the description under which you intended to do it – a claim that originates in Anscombe (1957). So whenever I intentionally form a belief that p via judging that p, this entails that I know what I am doing by the description under which I intend to do it. Hence, I know that I am forming a belief that p. Moreover, Roessler takes this to show that whenever I intentionally judge that p I am in a position to know that I believe that p, because I know that I am forming a belief that p.

The problem with this stronger proposal is that it only works for intentional mental actions, but most of our belief-formations are not intentional mental actions. His account therefore has difficulties to be applied generally. Moreover, while there is no difficulty of intentionally judging whether something is the case, it is unclear whether and how this proposal can avoid the problem of scope and be applied to a variety of mental state types. The weaker proposal of only requiring an empirical, contingent correlation between first-order process and second-order process is easier to generalize for all types of mental states and mental actions. It comes with the downside that it requires empirical work in addition to the discussion of structural features that I can provide here.

[65] As proposed by Boyle (2011).
[66] Though Peacocke (1998) has a related position.

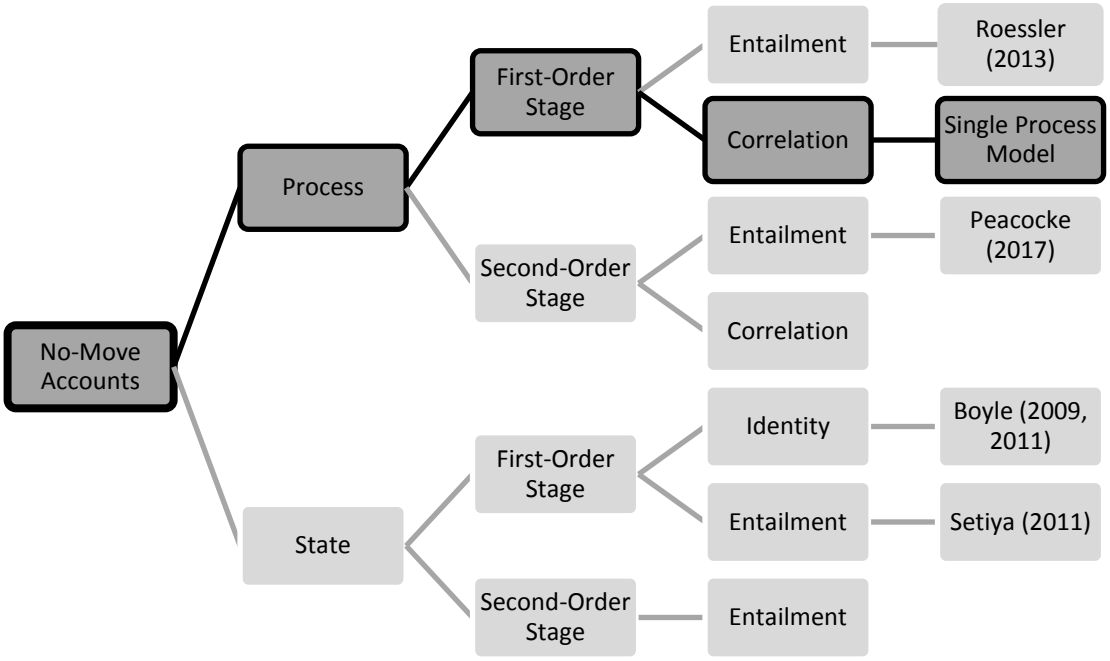

Figure 3 – The Single Process Model in the Taxonomy

4.4 The Single Process Model: The Account

With the general direction of the single process model outlined, I can now fill in the details of how the view explains self-knowledge. Whereas move accounts suppose that there is a specific transparency method that transitions one from the outward phenomenon to a self-ascription of a mental state, the single process model posits there is no such method whatsoever. Rather, whatever one does to find out whether one believes that p is the same procedure that one employs when one wants to find out whether p is true. Evans's points towards this reading when he claims that "I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (1982, p. 225).[67] Moreover, he reiterates the point two pages later about self-ascribing perceptual experience when he claims:

> However, a subject can gain knowledge of his internal informational states in a very simple way: by re-using precisely those skills of conceptualization that he uses to

[67] Silins reads this passage vastly different. He thinks that Evans emphasizes active "[…] opening enquiry into whether p […]," (Silins, 2013, pp. 295-296) instead of sameness of procedure for first- and second-order belief formation. Bar-On also does not read much into sameness of procedure (Bar-On, 2004).

> make judgements about the world. Here is how he can do it. He goes through exactly the same procedure as he would go through if he were trying to make a judgement about how it is at his place now, but excluding any knowledge he has of an *extraneous kind* (p. 227).

While self-ascribing perceptual experience does not work exactly like self-ascribing beliefs[68], it nevertheless fits the same idea of outward directed self-knowledge. Hence it is useful to find one overarching way of understanding transparency. I want to highlight Evans's[69] emphasis on "[…] precisely those skills […]" and "[…] exactly the same procedure […]." These formulations indicate that he did not have a special method of transparency in mind, which we can use to produce self-knowledge. Rather he took us to form self-ascriptions of mental states by invoking our usual skills and procedures.[70] Whatever we use to generate a mental state is also exactly the thing that is used for self-ascribing it. For the belief case this means that there is no process that we need to add to whatever we do when we form first-order beliefs. Everything is already present. However, for this to be true a single process has to be able to produce more than one belief. Moreover, the process needs to be able to generate beliefs of various orders – first and second. So my proposal in a nutshell is this: normal, adult human beings employ a single process that forms first-order attitudes and second-order beliefs. And because this process is responsible for both of these attitudes Evans's slogan of getting "[…] myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (Evans, 1982, p. 225) is satisfied. Moreover, because the procedure involves generating first-order attitudes about the world it also satisfies the idea of attending to an outward phenomenon.

Note that Evans writes about "re-using" and "exactly the same procedure". These two seem to clash. If it is exactly the same procedure then re-using it (after you used it before) should not give you any new beliefs. However, in a later section on transparency I show how to reconcile this in virtue of two senses of transparency. I thereby will depart from Evans to some extent. For now just take the idea of the same procedure as the central one.

---

[68] Perceptual experience relies on non-conceptual states according to Evans (1982, p. 227).

[69] At this point I need to note that I do not claim that this is exactly what Evans had in mind. Rather I want to explore the possibilities of a certain reading of his work.

[70] There is a sense in which Byrne's proposal also satisfies this, because the epistemic rule only uses processes that are used for first-order belief formation in other cases. However, Byrne still requires that one uses something in addition to first-order attitude-formation. In his case you just add something that can also be present in first-order cases (reasoning).

My proposal can be seen as a commitment to either the strong (The procedure is exactly the same as in forming the first-order mental state that is self-ascribed), or the weaker (the procedure is latched onto first-order mental state formation) reading of *sameness of procedure*. I choose the latter one. My rationale for this choice is the following: On one hand it gets the idea right that there is no distinct, independent belief-formation process of introspection. What we do when we self-ascribe a mental state is not distinct from what we do when we form first-order mental states. Generally, a first-order mental state and the corresponding second-order belief are formed in one sweep. On the other hand, by opting for the weaker reading the account opens options of how to explain mismatches between first-order mental state and second-order belief about one's mental states.[71] The strong reading assumes that there are no differences in mental state formation at all, so mismatches are difficult to explain. In contrast, the weaker reading proposes that even though the second-order belief is formed partially on the same grounds (as it is merely latched onto the first-order mental state formation), and is dependent on the process generating a first-order mental state, it can be influenced by factors that are outside the scope of the first-order mental state formation. Hence, when I am talking about 'a single process' this should be understood as a first-order process with a latched on second-order process. The second-order process cannot exist independently from the first-order process, and generally the first-order process comes with the second-order process.

I can now propose the following principle of the single process model (SP), wherein 'attitude' includes both propositional and non-propositional attitudes and the process P refers to a token process:

(SP) A single attitude-forming token process P produces both my first-order attitude A and my belief B that I have attitude A*. A* is identical to A, if everything goes right.

For instance, in a good case a single token process produces my desire to ϕ and my belief that I desire to ϕ. Because everything goes right, the first-order attitude produced and the content of the second-order belief match – both are a desire to ϕ. In a bad case these would not match. One might generate a desire to ϕ, and a belief that one desires to ψ.

[71] I discuss an independent motivation for this later.

This principle has to be clarified in a couple of aspects. First, the principle requires the subject to possess mental states concepts; otherwise the second-order belief would not be possible. Therefore, I assume a normal, adult human being as the subject. This leaves developmental and concept acquisition questions open.

Second, attitude forming processes should be understood in a very broad sense. These include, but are not limited to: forming a belief, a desire, a hope, a fear, a wish, etc.… Importantly it includes non-propositional attitudes. It is not entirely obvious which attitudes are non-propositional. I may fear snakes, which seems to be non-propositional, but on the other hand I may fear that it will be -20° C tomorrow morning, which seems to be propositional. Even desires, which are usually taken to be propositional, are contested to be non-propositional after all (Thagard, 2006). Moreover, one may be in general wary of describing mental states as propositional (Ben-Yami, 1997). I take Grzankowski's characterization of non-propositional attitudes to be generally correct: "[…] there are attitudes that relate individuals to non-propositional objects and do so not in virtue of relating them to propositions. Examples include loving, liking, hating and fearing, though there are probably many more" (2012, p. 1). For both propositional and non-propositional states I leave open what a process of attitude production actually is or includes, as I take that to be an empirical question answered by psychology. But for the purpose of this model it is enough to think of the attitude forming process of, say, visual perception[72] as a complex of the object, light, visual organ, neural activity, etc. The employment of attitude-forming processes helps to make the account applicable to all kinds of attitudes and not limited to self-ascription of belief.

Third, first-order attitude A and the belief B are not conceptually or constitutively connected. They are produced by a single process, but the attitude and the embedded attitude can differ. This opens the account to fallibilism. I may have an attitude A without the belief that I am in A. I may rather believe myself to be in a similar state A*. Think again of the classic case presented by Peacocke (1998): An administrator claims to believe that graduates of foreign universities are equally qualified as those graduates of her own country. However, in hiring decisions to be made solely on qualification, she gives preference to her compatriots. (SP) can explain this by accepting that the administrator has

[72] For an overview on the process of perception according to psychology see Bruce, Green & Georgeson (2003).

an attitude A (the belief that compatriots are preferable), but she does not have the corresponding second-order belief B to be in A. I will look into an explanation of this mismatch in a later section.

Fourth, the formulation in (SP) is quiet on the basis of the attitude forming process. Moreover, (SP) in its current form is quiet on what the process looks like. This opens up at least two different options. In one version both the first-order attitude and the second-order belief share a basis completely or partially. In another version their basis differs significantly, but usually both are generated as part of a single process. This difference is best understood as the difference between a bypass version and a monitoring version. The bypass version is committed to the following principle:

(Bypass) Neither attitude A, attitude A*, nor belief B are part of the attitude-forming process.

The bypass version takes attitude A, attitude A*, and belief B not as part of the process insofar as they are only products of, but not a basis for anything in the process.

The alternative to bypass is some kind of monitoring:

(Monitoring) Attitude A is the basis for forming belief B.

Monitoring is best understood by considering Nichols' and Stich's (2003) account.[73] Suppose you form a belief that p. Nichols and Stich think of this as 'p' going into a belief box – which represents a functionally characterized processing and storage mechanism. Believing some content just means having the content in the belief box. Now you could further suppose that generally when 'p' goes into the belief box it automatically gets copied, prefixed with 'I believe,' and the resulting 'I believe p' also goes into the belief box. This picture could also be described as an instance of (SP). As part of one process 'p' gets into the belief box, is copied, prefixed, and the prefixed copy also put into the belief box. Here your second-order belief is based on your first-order belief. Hence, your first-order belief is part of the single process as a basis for the second-order belief.

I have no knock-down argument against the monitoring version. However, there are some reasons in favor of bypass. Most importantly, the bypass version is simply more economic

[73] Thanks to Alex Byrne for suggesting this comparison with Nichols and Stich.

than the alternative. The monitoring version requires a way to generate a belief on the basis of an attitude. To do that one needs some way to pick out the attitude. In a picture inspired by Nichols and Stich this would be achieved by a monitoring mechanism that detects a content that is put into the belief box. The bypass idea can avoid the need for any such monitoring mechanism. In addition, it is a further question whether a monitoring version would still be in the spirit of transparency accounts. After all, Evans thought of transparency as an alternative to inner-sense views. Finally, one might ask whether the monitoring version actually involves a single process. From here on I am only going to talk about the bypass version of (SP). Hence I will include (Bypass) in further adjustments to (SP).

Fifth, an implication of (SP) is that for every first-order attitude that is formed one usually gets a second-order belief for free. If there is only one attitude-forming process generating two attitudes it should be difficult and rare to find one of these attitudes generated on its own. It is tempting to claim that it requires the impossibility of a belief formed without a second-order belief. However, I find this to be too strong. It seems possible that an attitude-forming process stops before generating any belief whatsoever. Perception just ends with a belief if everything functions properly. Similarly (SP) only requires that attitude-forming processes generate a first-order attitude and a second-order belief if everything functions properly. A bad case might, for instance, be one in which the subject cannot employ mental state concepts and therefore produces a first-order attitude without a second-order belief.

That first-order attitudes generally go with a corresponding second-order belief may still seem troublesome. This seems to imply that generally when one forms a first-order attitude, one also becomes aware of this attitude. An obvious mistake, considering one certainly can be unable to report one's attitudes. For instance, if I was asked at the moment I am writing this, what mental states I am in, I could not give an answer.[74] However, I can remedy the problem by accepting that the production and the access[75] of the second-order belief can diverge. One way to spell this out uses the notion of *dispositional* belief.[76] This

[74] This is also pointed out by Schwitzgebel (2008, p. 251).

[75] Access is to be understood in a very undemanding sense. Access to a mental state need not be consciousness of that state. Accessing a mental state is simply for that mental state to become occurrent.

[76] It is important not to mix up 'dispositional belief' with 'disposition to belief.' A dispositional belief involves a mental state with some particular content stored in the brain, and this mental state can

notion is best introduced by a simple demonstration: Think of your birthday. Certainly, you knew when your birthday is before you thought about it consciously right now. Nevertheless, only when I asked you to attend to it your knowledge and thereby your belief, was *occurrent*. Before that moment you only had the dispositional belief. Only in the right circumstances this dispositional belief becomes occurrent. Spelling out dispositional belief is tricky, but for my purpose it should be sufficient to say that *S* dispositionally believes that p iff[77]

(i) S has endorsed the content of p;
(ii) S has stored this content;
(iii) S can recall this content in the right circumstances; and
(iv) the content of p affects S's behavior, reasoning and mental states in the right circumstances

Whereas an occurrent belief is a belief that is endorsed and currently recalled or active corresponding to (iii) and (iv).

Endorsement should be understood in a way that renders it undemanding and does not require any consciousness whatsoever. This is important, so that dispositional believing does not require anything close to prior occurrent believing, which safes the criteria from becoming too intellectualist. The notion of dispositional belief used here is not particularly idiosyncratic. Robert Audi already explicitly states that "*the occurrence of belief formation* apparently does not entail that of an *occurrent* belief" (1994, p. 420), even though he takes forming dispositional beliefs without prior occurrent beliefs to be rare. Nothing requires a dispositional belief to have been occurrent before. I find it helpful to illustrate dispositional beliefs further with Audi's computer analogy. We might think of dispositional beliefs as information stored in the computer's memory, but not shown on the screen. In some circumstances the information from the memory will be read and then shown on screen. This is similar to the dispositional belief becoming occurrent. In this picture, a dispositional belief being formed without any prior occurrent belief can be simply illustrated as some information being stored in the memory without ever showing up on screen. There is nothing inconsistent about this possibility (Audi, 1994, pp. 420-421).

---

become occurrent in the right circumstance. On the other hand, a disposition to belief is merely a disposition to form a mental state with some particular content when one is in the right circumstances. For a discussion of dispositional belief see Audi (1994).

[77] These rough and ready conditions are inspired by Gertler (2011b) but differ slightly.

I want to emphasize that my proposed definition of a dispositional belief does not identify occurrent beliefs with conscious beliefs.[78] The difference I want to pick out is rather the active role the belief plays. An occurrent belief is active in some sense, whereas the dispositional belief is stored, but currently not doing anything. An occurrent belief can be active insofar as it is conscious, but it can also be active without being conscious, e.g. when it affects the behavior.

We now get the second version of the single-process principle:

(SPD) A single attitude forming token process P produces both my first-order attitude A and my dispositional belief B that I have attitude A*. A* is identical to A, if everything goes right. Neither A, A*, nor B are part of the attitude-forming process.

This version handles the problem of having awareness of all of one's propositional attitudes. While generally propositional attitude come with second-order beliefs, these beliefs are neither permanently occurrent nor accessible. Only in the right circumstances one's beliefs about one's propositional attitudes become occurrent. This may be a case of triggering the occurrence with a question or prompt (e.g. asking you to think about your birthday). Hence, (SPD) can be used to explain successful, transparent self-ascription in two stages. First, the belief forming stage: I perceive a tree in front of me and thereby form the belief that there is a tree and the dispositional second-order belief that I believe there is a tree. Second, the stage at which the belief becomes occurrent: A friend of mine asks me whether I believe there is a tree in front of me. My dispositional belief gets triggered by this question and becomes occurrent. I answer that I do believe there is a tree in front of me. Both stages can coincide.

There is a sense in which the first stage alone already involves self-ascription, because it involves the dispositional second-order belief. However, it is possible that some such self-ascriptions never become occurrent in the actual world.[79] In our linguistic practice we trivially cannot avow these self-ascriptions, because avowing requires occurrent belief. For this reason only those second-order beliefs that have been occurrent at some point or are

[78] For reasons why we should not identify occurrent beliefs with conscious beliefs see Bartlett (2017).

[79] They still have be occurrent in some possible worlds to satisfy (iii) and (iv) of the conditions for dispositional belief.

occurrent now are denoted by our ordinary talk of self-knowledge. These are self-ascriptions in a *narrow* sense.

If (SPD) is correct then we constantly form dispositional second-order beliefs. One may be skeptical of this based on cognitive limitations. How can we store all these beliefs? Even though this question sounds legitimate at first, I do not think this is a heavy hitting problem. We already accept that we have a huge amount of dispositional beliefs anyway. Think again of the birthday example. There are a vast number of similar cases of propositions that we dispositionally believe at any given time. I find no good reason to why accepting more dispositional beliefs is problematic.

Sixth, I claimed that A* is identical to A if everything goes right, but I need to elaborate on this condition. Self-knowledge is only possible if A* is identical to A, because only then truth is guaranteed. However, a second-order belief might still be true by luck even if A* is not identical to A. Consider an implausibly scared person Jack. Jack is scared of every animal he perceives, or ever thinks about (both real and imagined). One day Jack sees a bear and as expected he is scared. Unfortunately, his second-order belief forming process goes awry and does not produce the belief that he is scared of a bear. Rather he comes to believe that he is scared of a bird. This belief is true, because Jack is scared of everything that crosses his mind and he just thought about the bird embedded in his belief.[80]

However, why should the second-order belief production bring about a belief about the attitude A* instead of A? I already pointed to the explanatory advantages to cope with cases self-deception as presented in Peacocke (1998), but I did not explain my thoughts about the mechanism that may lead to false, or only luckily true second-order beliefs. I find it plausible to assume that beliefs are not formed in vacuum. Rather, I take it, other mental states, especially other beliefs, can influence the belief production. That is, some cognitive penetration exists for most, or perhaps all belief-formation.[81] Some (not fully conclusive) evidence that points to this influence is presented by Churchland (1988; 1989), Balcetis & Dunning (2006), Stefanucci & Proffitt (2008; 2009), and van Ulzen et al. (2008). Given

[80] This is similar in structure to Gettier-style cases for fallibility in general, as discussed by Baron Reed (2002)

[81] For an overview on cognitive penetrability of perception see Stokes (2013). Debates on cognitive penetrability usually focus on perception. This happens because perception at first sight seems independent of standing mental states of the subject. So if perception turns out to be influenced by other states, so will likely other belief-forming processes.

cognitive penetrability we can expect that first- and second-order beliefs can be differently influenced by other mental states. So while it is a single process that produces first-order attitude and second-order belief, they do not need to be formed exactly the same way. Second-order beliefs can be differently affected by other mental states than first-order attitudes. Think of the process as follows:[82]

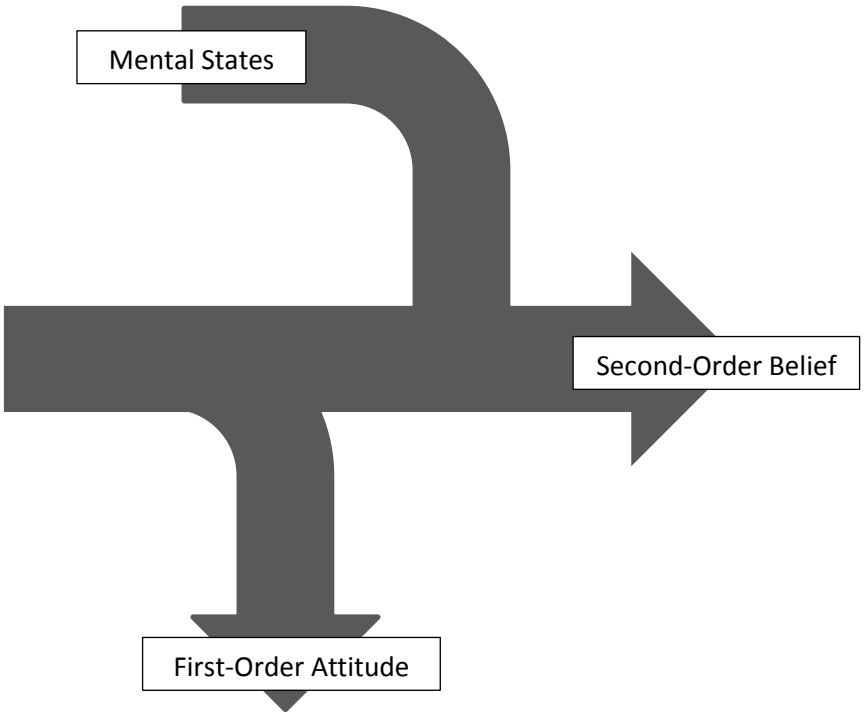

Figure 4 – Mismatches in the Single Process Model

The process produces the first-order attitude and the second-order belief, but the second-order belief is influenced differently by other mental states. First-order attitude and second-order belief share large parts of their origin which accounts for reliability. In other words, it explains A* being identical to A if things go right. The second-order belief, however, is prone to additional influence of other mental states. This brings us to the final version of the single process model:

[82] This shows a case in which the second-order belief is influenced by mental states that the first-order belief is not. I thereby do not claim that first-order attitudes are not influenced by other mental states, even though there is no other input arrow in the graphic.

(SPDM) A single attitude forming token process P produces both my first-order attitude A and my dispositional belief B that I have attitude A*. The production of B can be influenced by mental states M, which accounts for the possible difference of A* and A. A* is identical to A, if everything goes right. Neither A, A*, nor B are part of the attitude-forming process.

A story of successful self-knowledge in this model looks like this:

> You look with full awareness, and under normal conditions, at a nearby red car. The process of perception causes you to believe that there is a red car. The process also produces the second-order belief that you believe there is a red car. It is, in a sense, correct to claim that perception produced this second-order belief, and that it was caused by a complex of whatever is part of the process: the car, light, visual organ, neural activity, etc.

I can further define necessary and sufficient conditions for self-knowledge of attitudes based on (SPDM).

> S knows her own attitude A based on a single-process iff:

1. S has a belief that S is in A
2. S's belief that S is in A is produced by the same token process as A
3. S's belief that S is in A is reliably produced
4. S is in A[83]
5. S has no relevant undefeated defeaters

We may also say that S can correctly avow her own attitude A only if furthermore[84]

6. S's belief that S is in A is occurrent

The conditions here match the standard account of knowledge as justified true belief plus an anti-luck condition. (1) accounts for belief, (2) and (3) for justification[85], (4) for truth. (5)

[83] One might think that this is already presupposed by (2). However, it seems plausible that (2) is satisfied without (4). Consider a case in which I form the belief that p and the belief that I believe that p in a single process. Then later I lose my first-order belief, but not the second-order belief. If (4) was not its own condition I would count as knowing my own attitude. But clearly I am not, because I do not have the first-order belief anymore.

[84] This is a necessary but not sufficient condition for avowing A. There can be cases in which a belief is occurrent and one is still unable to avow that belief, e.g. when a belief affects behavior unconsciously.

[85] Given a reliabilist notion of justification as provided by Goldman (2008 (1979)).

is a general defeater clause. Self-knowledge is achieved just in case the process produces both my propositional attitude A and the dispositional belief that I am in A. Moreover, the process has to be reliable with regard to my second-order beliefs. Cases of veritic luck (such as the implausibly scared Jack above) are already ruled out with (1) to (5), so I do not require an additional anti-luck condition. With this set in place I can explain key features of self-knowledge. I link the single process model to epistemic asymmetry, reliability, and fallibility. Moreover, I elucidate the claim that the account qualifies as a transparency account.

4.4.1 Asymmetry

I can employ the most natural and simple explanation. Self-knowledge seems to be different from knowledge of others' attitudes, because it is different. Knowing the mental states of other people requires one to observe and interpret their behavior. Fernández (2013) is right that my way to form a belief about him wanting Barcelona FC to win the UEFA Champions League is by hearing him expressing that desire, watching him cheer for the team or be sad when a defeat is reported. My justification relies on behavioral evidence and reasoning. When I believe that I want AFC Ajax to win the UEFA Champions League, I do not rely on observing my own actions and then interpreting the evidence. Rather I have this belief immediately as it was produced with the formation of the desire for AFC Ajax to win. And now I just access this second-order belief by thinking about the Champions League. Neither the stage of belief production, nor the dispositional belief becoming occurrent, requires any interpretation of my behavior. However, this is not due to some constitutive relation of the self-ascriptions to first-order attitudes, but rather due to the empirical circumstance that one's attitude-forming process happens to also generate dispositional second-order beliefs. And this takes place largely in one's head, which explains why the same thing does not work for other people's mental states. Their attitude-forming process cannot generate a second-order belief in me, because I am not connected to other people's attitude-forming processes the same way I am connected to my own. This is primarily a question of location.[86] If we could be interconnected in the right way it is at least metaphysically possible that a single process would generate a first-order attitude in person X and at the same time a second-order belief in person Y.

[86] Thanks to Miriam McCormick for this formulation.

The single-process model also provides an explanation of the phenomenology of self-knowledge. In chapter 3 I raised the problem that only in some cases of self-knowledge the belief-formation is experienced as attending to an outward phenomenon, and hence experienced as formed transparently. The prime example for this is Evans's case of being asked whether one thinks there will be a third world war. However, in a majority of cases self-knowledge appears to be immediate without any awareness of attending to an outward phenomenon. In these cases self-belief seems to be formed 'silently' (O'Shaughnessy, 2000). The single process model can account for this. There is no distinct process of forming second-order beliefs, but there can be a distinct stage of retrieving second-order beliefs in which they become occurrent. Second-order beliefs are formed together with the first-order attitude. Generally, if you have a first-order attitude, you also have the corresponding second-order belief. Adding the self-ascription to the first-order belief-forming process allows a distinction between two different types of cases of avowing one's attitude:

a) Cases in which first-order attitude and second-order beliefs were formed right before the avowal; and
b) Cases in which first-order attitude and second-order beliefs were not formed right before the avowal.

These two types of cases correspond to the phenomenological observation: In cases of type (a) the phenomenology fits naturally with the transparency claim. Evan's third world war example is, usually, a case best described as (a). If someone asks me 'Do you think there is going to be a third world war?', I (most likely) will not have formed any opinion on whether there is going to be a third world war. Hence, I must attend to the question 'Will there be a third world war?' and in doing so will form both a first-order belief about a third world war and a corresponding second-order belief. I can then avow whether I believe that there will be a third world war. Because the formation of both beliefs happens right before avowing, the phenomenology of acquiring self-knowledge fits transparency here.

In cases of type (b) the phenomenology fits the silently formed, immediate self-ascriptions. Hence, if someone asks me whether I believe that we are in the year 2018[87] I can answer immediately. The case is best described as (b). I can answer that I believe us to be in 2018

[87] A reminder: You might not be convinced by this example because 'do you believe that x?' questions could be interpreted as merely different ways to express 'is x the case?' questions. However, the same immediacy of answering questions about my mental states can occur when one is asked whether one desires x/desires to Φ, wants x/wants to Φ, intends to Φ, etc..

instantly, because I formed a second-order belief that I believe it to be 2018 sometime in the past and stored it. Now my stored second-order belief merely has to be retrieved from memory and become occurrent. This is an easy and quick process, which explains the experience of immediate self-ascription of the belief silently. It still has been generated according to transparency, but this generation happened in the past when the first-order belief was formed. Hence, the single process model provides a plausible story of how our phenomenology of forming self-beliefs comes about, and why it differs from the phenomenology that comes with forming beliefs about other people's mental states.

4.4.2 Reliability

Reliability is explained by the close connection between first-order attitude and second-order belief. They are produced by a single process and share most of their origin. They have, if everything goes right, the same basis. Nevertheless, this is not a guarantee. There is no constitutive relation securing that first-order attitude and second-order belief match. That they usually do is a contingent matter. We just happen[88] to be wired in a way that we generate first-order attitude and dispositional second-order belief by a single process such that they usually match. Consider the red car example from above. The perception process generated the belief to see a red car and the corresponding second-order belief. And they reliably match because generally when I form the second-order belief that I'm in attitude A by this process I also form the attitude A. That whatever is responsible for the first-order attitude also determines the second-order belief plausibly plays a role in establishing the reliable correlation, but I'll remain neutral on the precise mechanism at work. As I mentioned in the beginning, this is not completely satisfying, but this is as far as the single process model takes us without scientific backup.

Which mechanism fits best is largely a question for empirical sciences. The single process model can remain neutral[89] on the exact mechanism while still correctly predicting and explaining characteristic features of self-knowledge. I only have to suppose that there is some such mechanism. Therefore it is sufficient that as long as nothing interferes in the middle of the attitude-forming process the second-order belief will be reliably the correct one for the first-order attitude. And this holds not only for belief, but for all attitudes.

[88] At this point one may wonder why we should be wired in this specific way. I sketch an answer with an evolutionary just-so story at the end of this chapter.

[89] Though not completely neutral. Some detectivist methods are ruled out in virtue of committing to the second-order belief bypassing the first-order attitude.

4.4.3 Fallibility

That you can fail to correctly believe, or assert what mental state you are in is established in two different ways. Either something interferes with the belief-forming process, or something interferes with your dispositional belief becoming occurrent.

In the first case another attitude may provide an incentive towards holding a certain belief. To illustrate this consider this case of an interfering desire:

> You wait in a restaurant for a friend whom you think you are supposed to meet at 6:00 p.m. It is already 6:30 p.m., and your friend is normally on time. You believe that you have the correct time and date for the rendezvous, but your friend is nowhere to be seen. After another ten minutes, your belief dissolves into mere hope. Nevertheless, your desire to see your friend is so strong, that for a while you still think and sincerely tell yourself that you believe you are correct about the scheduled date and time in question. You are wrong about yourself, because there is already only hope left and no longer belief. Another thirty minutes later, you leave with disappointment.[90]

Slowly you lose your belief and it is replaced by an attitude of hope. However, your desire interferes with the second-order belief production. While you ought to form a belief that you hope that your friend will arrive, you hold on to the belief that you believe that your friend will arrive instead.

The second way to false avowals opens up when a true dispositional second-order belief is produced, but it cannot become occurrent. Take a therapy case that is often regarded as a case of self-knowledge being interpretational (this is a slightly modified case taken from Moran (2001, p. 85)).

> Carla feels anger at her recently deceased parents for having abandoned her. However, this is a suppressed attitude. She is easily irritated and has violent outbursts but does not have any conscious thoughts about this anger being linked to her parents. Carla goes to a therapist. After talking to a therapist the evidence laid out by her and the therapist convinces her that she actually feels this deep anger targeted at her dead parents.

[90] The example is inspired by Bernecker (2010, p. 232).

There is a certain intuitive appeal to describe what is going on in therapy as interpretational self-knowledge. There is obviously some sort of interpretation going on. Nevertheless, I think this is not necessarily part of forming the second-order belief. I described self-knowledge as a two stage process: First you form the dispositional second-order belief together with the generation of the first-order attitude. Then, as the second step, you have access to this second-order belief in the right circumstances. Frequently this is already achieved via the situation that made you produce the attitudes in first place. Sometimes, however, the access to the second-order belief is more difficult. Complex interactions of psychological states hinder you from coming to an occurring belief about your own attitude. At this point therapy may intervene. So when Carla comes to know that she feels anger at the dead parents this can be explained as a case of non-interpretative self-ascription. The interpretation certainly leads to occurrent self-knowledge, but it does not produce this self-knowledge. The interpretational work provides circumstances in which Carla has occurrent self-knowledge. The therapy does not produce the second-order belief; it only provides a context in which Carla has access to her dispositional belief, which was already present before. Without therapy there would not be this context and she would lack the occurrent belief necessary to avow that she is angry at her parents.[91]

The difference between dispositional and occurrent belief can also be used to provide a similar story for commissurotomy ("split-brain") patients. Consider the following experimental setting (Gazzaniga, 1995; 2000; Carruthers, 2010; 2011): A subject's left and right parts of the brain have no connection to each other. An instruction "walk" is flashed to the right side of the brain and the subject begins walking. When the subject is asked to explain why he was acting the way he was acting, his left-side brain, which is in control of speech-production, produced the answer "I'm going to get a Coke from the house". Carruthers claims that this is completely confabulated, but nevertheless the subject answered confidently and immediately.[92] Even when the subject is reminded of his

[91] This is not a claim that therapy cannot generate second-order beliefs by interpretational means. Rather, I use a particular therapy case to present a source of false avowals that is available with the single process mode. Unfortunately, I have not yet come up with a clear case that has nothing to do with therapy.

[92] Carruthers (2010) states explicitly that the subject confabulates the reason why they stood up. However, he concedes that it might be the case that the subject would have actually gotten a coke if the experimenters would not have stopped him. Carruthers explains this as the confabulation becoming self-fulfilling afterwards. In his 2011 book this self-fulfillingness is left out and the question changed from being asked *why* the subject is walking to *where* the subject is going. This difference matters, because only the why question seems to indicate a failure of self-knowledge. Gazzaniga's

condition the confabulations continue. And it seems to be phenomenally just like ordinary introspection to the subject (Carruthers, 2011; 2010). Suppose that Carruthers interpretation of the case is correct and it actually shows a failure to generate self-knowledge. Can the single process model explain what goes wrong?

The single process model can give a principled story for this systematic confabulation. (SPDM) predicts that when the subject reads the "walk" instruction the subject forms the intention to walk around according to the instruction, and the subject also forms a corresponding, dispositional second-order belief. So the subject has a dispositional belief that she intends to walk around according to the instruction. However, this dispositional belief is stored in the side of the brain that the card was flashed to. When the experimenter later asks for the reasons for walking, the experimental setting together with the subject's particular brain architecture makes sure that the other half of the brain is tasked. This implies that the dispositional belief cannot become occurrent whatsoever, because it is not connected to the now tasked part of the brain. Having no access to their second-order belief whatsoever causes the subject to confabulate when asked. This confabulation can constitute a new second-order belief B* formed on weak grounds. This second-order belief then is likely unreliable and hence usually false.

The possibility that a dispositional belief does not become occurrent presents a potential worry for the single process model. In chapter 1 I discussed the apparent authority of avowals as a phenomenon that should be explained by accounts of self-knowledge. The single process model explains the apparent authority in virtue of reliability: Avowals are authoritative because they have a good chance of being true. However, because avowals require occurrent belief this requires the process of the dispositional belief becoming occurrent to be easy. If there were frequent issues with stored beliefs becoming occurrent the apparent authority could not be explained. The mere fact that the dispositional, second-order beliefs are reliably formed would not be enough, because the potential failure for them to become occurrent could make the avowals lack the required preponderance of truths over falsehoods. The single process model therefore has to presuppose that occurrent beliefs can become occurrent relatively easily. The stronger claim that they can always become occurrent is not needed, but at least most second-order

(1995; 2000) discussion includes experimental setups with why-questions, so Carruthers' interpretation cannot be ruled out based on this issue.

beliefs about one's mental states should become occurrent when triggered by a question about whether one has a particular mental state. Usually, being asked whether one has attitude A is enough to trigger the corresponding dispositional belief to become occurrent.

Explaining the apparent authority of avowals poses another potential problem here. The moment the second-order belief is formed and then stored as a dispositional belief can be far apart from the moment one actually avows having a mental state. There is some risk that over time your dispositional belief is not perfectly retained, and hence the high likelihood of the belief being true when formed initially might not be preserved. Suppose one becomes angry with Boris and at the same time forms the dispositional belief that one is angry with Boris at t_1. Nothing changes with this anger in the next years. Five years later one is asked whether one is angry with Boris. In a good case this would trigger the dispositional belief into becoming occurrent. However, because the second-order belief had to be stored for so long that belief might not be preserved correctly. Even though consolidated semantic memory[93] is quite stable, it does not retain information perfectly, so it is possible that over the years something got lost, or changed. Given this possibility, some of the epistemic goodness that the second-order belief had when it was formed might be lost as well. The idea that dispositional beliefs can become occurrent easily will not help here. If the dispositional belief was changed while stored the ease of occurrence will not change its probability of being true. Moreover, even if the second-order belief was stored properly, it might have not been maintained properly in a different way. The second-order state has to be minimally sensitive to changes in one's mental life, such that if the first-order anger would cease to exist, the corresponding second-order belief has to be discarded as well. Usually this is no problem, because according to the single process model changes in first-order states correlate with changes in second-order states such that they reliably match. However, this opens up a further way of avowing falsely when the second-order beliefs are not appropriately updated with changes in the first-order states. Given that these error possibilities increase with the duration of the storage of second-order beliefs the single process model has to concede that a subject might not be in a privileged position in avowing mental states that she formed a very long time ago. With time passing between the second-order belief formation and the belief becoming occurrent, the probability of the avowal's truth decreases. In some cases the avower might not even be in

[93] Memory consolidation refers to a process that stabilizes a memory trace for long-term storage. For an overview of research on consolidation see Dudai (2004).

a privileged position to self-ascribe her own mental states because the dispositional belief is too likely to have changed in storage.

So why do we then still take avowals to be authoritative, according to the single process model? My answer is twofold. First, most of our avowals are based on dispositional beliefs that are either formed recently, or have been occurrent recently. More recently formed dispositional beliefs have a better chance of being retrieved without any problems or changes to the content. Furthermore, anytime a belief becomes occurrent we are in a position to reevaluate[94] the belief in a way that can strengthen the memory trace. This fits with research on reconsolidation (Cf. Sara (2000), Rodriguez-Ortiz & Bermúdez-Rattoni (2007)) that observes that memory traces are especially malleable in moments of retrieval and links that fact to better long-term storage. Second, the hearer of an avowal is not in a position to know when the avowed mental state was formed initially or how the stored belief was maintained. Sometimes the content of the avowal will provide some information about the origin of the mental state (e.g. being angry with Boris when the hearer knows that the last time the speaker interacted with Boris has been a long time ago), but usually the hearer will have to treat the avowal as based on a recently formed mental state simply because it is the most probable option. And because avowals of recently formed mental states are likely to be true, the avowal will be taken as authoritative as long as the hearer has no reasons not to. However, if the hearer has reasons to believe that the avowed mental state has been formed a long time ago they should not treat the avowal as authoritative anymore. This fits with our ordinary linguistic practice and explains why at some point an interpretational second-order belief formation (self-directed mind-reading) might be the best option. Go back to the case above: After a long time you might still believe that you are angry with Boris and if asked say so. But it does not seem inappropriate for a friend to ask you again if you really still feel anger or whether it might be something else. Moreover, we can imagine a plausible case then continuing with your reassessment of the anger until you find out that you are not actually angry anymore, but, say, merely disappointed and hostile towards Boris. This can be interpreted as a case of improperly maintained second-order belief. At some point one's first-order state changed, but the corresponding second-order belief did not. The linguistic practice shows some sensitivity to the probable recency of the mental state an avowal is based on and thereby is also

[94] This process will be unconscious most of the time.

sensitive to the probability of the avowal being true. Hence, here the linguistic practice seems not to be a problem, but rather an ally for the single process model.

4.4.4 Transparency

I promised a transparency account of self-knowledge. This is satisfied in two different senses. Similar to fallibility being explained at the level of belief production and at the level of belief access, transparency can be found in both levels as well. First, let me consider the belief production side of things. Remember Evans's formulation of transparency:

> If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p (Evans, 1982, p. 225).

When the first-order attitude and the second-order belief are produced by a single process this is in a sense trivially satisfied. Consider the case of looking at a red car in front of me. The perception process, including all the outward parts, produces the second-order belief that I believe there to be a red car. It also forms the first-order attitude, the belief that there is a red car. Therefore it is true that I did put into operation the mechanisms that also generate the answer to the question whether there is a red car in front of me.

This schema also works for the original case. When I form a belief about whether there will be a third world war, I also form a corresponding, dispositional second-order belief. I can answer the question afterwards because my second-order belief becomes occurrent. The whole belief-forming process in this case is more complex. I weigh reasons for the likelihood of another world war to form my first-order belief. However, there is no reason to think that this more complex deliberation is in principle any different with regard to second-order belief formation.

There is a second way transparency comes into play. Whenever someone tries to find out whether she believes that p by putting into operation whatever procedure she has for determining whether p, she thereby creates a specific context of inquiry. This context may enable her to access a dispositional second-order belief. This is a way to capture Evans's idea of "[…] re-using precisely those skills of conceptualization that he uses to make judgements about the world. […]" (Evans, 1982, p. 227) that is compatible with the single process model. An example goes as follows:

> Aimee thoughtfully considered whether there will be a third world war at t_1. She formed the belief that there will be a third world war and the corresponding second-order belief at t_1. At t_2 she focused her attention on an article of the New Yorker and stopped thinking about the third world war. At t_3 she is asked whether she believes that there will be a third world war. Initially, the topic seems familiar to her, but for some reason the dispositional belief that she believes there will be a third world war does not become occurrent. So she starts thinking again whether a third world war will occur. In the process of considering relevant factors and before she comes to any conclusion suddenly the belief that she believes that there will be a third world war pops up.

By thinking about the possibility of a third world war Aimee created circumstances in which her dispositional second-order belief became occurrent. It is pivotal that in this case she only created an environment that gave her access to the second-order belief already in place. She did not generate a new belief, because the old one became occurrent before she concluded anything. Considering reasons for and against a third world war happening triggered the dispositional belief to become occurrent.

4.5 Advantages of the Single Process Model

In the previous section I showed how the single process model explains the features of self-knowledge I set up as the target phenomenon in the beginning. However, the single process might not be the only proposal that can explain these features, so why should we pick the single process model over other competitors? I discuss three advantages of my proposed model. First, it is economic, second it fits with phenomenological observations of transparency and immediacy, and third it avoids common problems for transparency accounts.

The single process model is economic insofar that it only relies on cognitive processes that are independently accepted anyway. The only processes required are attitude forming processes that one ought to be committed to simply in virtue of recognizing these attitudes as existing. If these attitudes exist, they have to be formed by some processes. The single process model only requires that these processes are also involved in generating self-beliefs. There is no additional, special kind of belief-formation necessary. As such the model is more economic than inner sense theories of self-knowledge, which require a special cognitive process of introspection. The single process model is not the only available

account with this virtue of being economic. For instance, self-other parity accounts such as proposed by Carruthers (2011) are equally economic. Therefore, this explanatory virtue alone is not sufficient for the proposed model to come out ahead.

The second advantage is a phenomenological one. As shown in part 4.4 the single process model vindicates both Evans's (1982) transparency observation and the phenomenology of silently formed self-knowledge (O'Shaughnessy, 2000). Hence, to the extent that these observations are phenomenologically plausible the single process model itself is phenomenologically plausible.

Finally, the single process model avoids common problems of transparency accounts. The current formulation works for all propositional and non-propositional attitudes. Hence, it avoids difficulties of scope that other transparency proposals struggle with (cf. Gallois (1996), Finkelstein (2012; 2003), Ashwell (2013a), Cassam (2014) and Chapter 3 of this thesis). Moreover, the single process model can account for self-knowledge of previously stored beliefs, which is also one of the common objections to traditional transparency accounts (cf. Shah & Velleman (2005), Gertler (2011a), Cassam (2014) and Chapter 3 of this thesis). The single process model can explain these simply by pointing to the stored second-order beliefs that were formed together with the first-order attitudes. I can know my previously stored beliefs in virtue of my second-order beliefs about these stored beliefs becoming occurrent. Both of these common problems are avoided in virtue of the single process model being a no-move account via linked processes at the first-order stage.

4.6 One Process or Two Processes?

Up to now I put the model in terms of a single process that produces two different outputs.

(SPDM) A single attitude forming token process P produces both my first-order attitude A and my dispositional belief B that I have attitude A*. The production of B can be influenced by mental states M, which accounts for the possible difference of A* and A. A* is identical to A, if everything goes right. Neither A, A*, nor B are part of the attitude-forming process.

Describing it as a single process has an advantage in pumping the intuition in favor of reliable introspection. When both first-order attitude and dispositional second-order belief are produced by a single process their close connection becomes very apparent. However, one might object that this is misleading and only tricks one into accepting these attitudes

being so tight-knit. The suggestion could look like this: What really happens is that two distinct belief production processes occur simultaneously. They may share some mechanisms and faculties, but they are distinct.

I agree that the same account can be modeled with two processes that usually occur together, but I deny that this entails any significant difference. Here is how a dual process account can look like:

> (First Process) A first-order attitude A is generated in an attitude forming process P_1. For instance: perception, desire-formation, production of a hope,…
>
> (Second Process) A second-order belief B that I have attitude A* is generated in an attitude forming process P_2. P_2 includes but is not limited to the mechanisms, faculties and causal relations of P_1. P_2 may further involve additional mental states M, regardless whether mental states are involved in producing A. A* is identical to A, if everything goes right. Neither A, A*, nor B are part of the attitude-forming process.

A story of successful self-knowledge in this model looks like this:

> You look with full awareness, and under normal conditions, at a nearby red car. The process of perception causes you to believe that there is a red car. A second process occurs simultaneously, which produces the second-order belief that you believe there is a red car. Both processes are caused by a complex of whatever is part of the process of perceptions: the car, light, visual organ, neural activity, etc.

There is no need for detection of any kind, no looking within or self-scanning. You just have the second-order belief in place and accessible by perceiving the car. Moreover it is true and justified, because your attitude production worked correctly. This is the case because I stipulated that no other standing mental states influence the second-order belief formation, and A* is identical to A. But one could imagine a case in which the second process produces a false belief because it is interfered by beliefs or desires as in the cases I discussed under fallibility. Figure 5 provides a sketch of the dual process model:

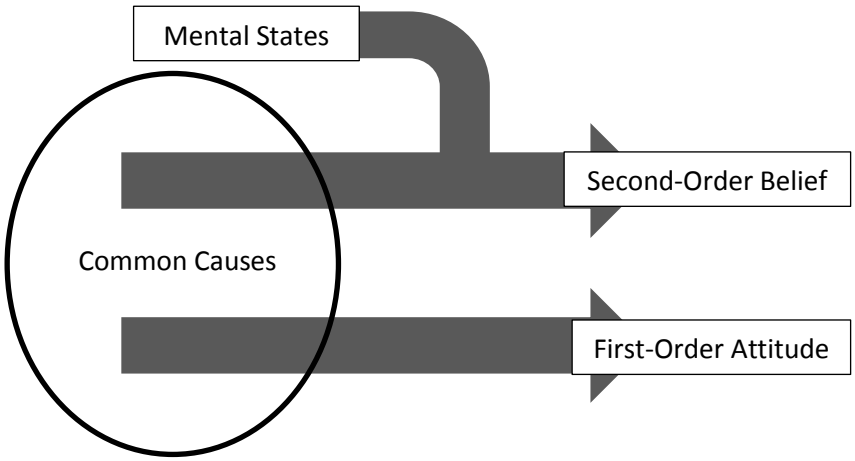

Figure 5 – The Dual Process Model

The dual process model can account for epistemic asymmetry, reliability, fallibility and transparency just like the single process model. Reliability and transparency are explained slightly differently though. Reliability is not established in virtue of first-order attitude and second-order belief being produced by a single process, but rather by common causes. Whatever causes a first-order attitude to be formed also causes the dispositional second-order belief. That both attitudes correspond is an empirical matter related to common causes and provides reliability.

Similarly, transparency cannot be accounted for by claiming that the very same process that answers whether p is also used to find out whether I believe that p, because this model uses two distinct processes. However, one may nevertheless claim that the same, shared causal mechanisms and faculties are in play. So there is a sense in which even under the two process model I acquire self-knowledge regarding my belief that p by putting into operation whatever procedure I have for answering the question whether p. For finding out whether I believe there is a red car in front of me, I still look at the car, use my eyes and rely on the neural activity taking place.

Transparency, in the sense of using the operations involved in first-order attitude production to create a context in which the dispositional second-order belief can be accessed, works just as it did in the single process version.

One motivation for using two processes is to have more resources available to deal with generality[95] concerns.[96] It is arguably easier (but nevertheless still difficult – and I am not offering any solution to the generality problem) to individuate processes adequately based on the generated state if every process only produces a single state. However, nothing stops one from treating the single process model as if it were two processes when considering reliability. Individuating processes can be done from various points of view and with various aims in mind. If we want to find out whether a specific mental state, a belief, is reliable we individuate the process that generated this belief based on the existing belief. And nothing stops us from doing so here. The second-order belief is reliable if the belief-forming process generates a good enough ratio of true beliefs. And this belief-forming process does not involve the first-order attitude because I committed to a bypass model.

However, if we focus on the processes producing mental states we can individuate the processes differently – from the other end, so to speak. Whereas individuating beliefs for reliability concerns starts with the result, we can start at the beginning of the process. We can individuate mental state production by looking at what goes on in a normal case of, say, perception.[97] What processes occur when one sees something?[98] (SPDM) has it that what happens in this case can be described as a single process producing two mental states.

Therefore whether we explain the process as a single process or two processes sharing causes is rather a matter of description. It is the question of where to begin our description – at the two distinct resulting states, or at the single shared starting point.

4.7 Challenges

I showed how the single process model explains features of self-knowledge and handles ordinary cases. In this final section I consider cases that create worries for the view.

The first challenge is motivated by a case of forgetting. Imagine Hazel forming the desire to be on TV. She also forms the belief that she has this desire. 8 years later she still has the desire, but she forgot the corresponding second-order belief. Hazel now wonders whether she desires to be on TV. Intuitively she can easily discover whether she does. But that

[95] For a formulation of the generality problem for reliabilism see Conee & Feldman (1998).
[96] This issue has been brought up by René van Woudenberg.
[97] You may substitute perception with any other first-order attitude forming process.
[98] Granted, there is a lot going on when one sees something that is irrelevant. However, by contrasting multiple cases of perception and other activities we can infer the specifics for perception.

cannot be in virtue of the single process account, given that she lost the second-order belief. If so, then why do we need the single process method at all?[99]

My response here is twofold. First, I question whether this is actually as straightforward as it seems. I have doubts whether forgetting the second-order belief while still easily discovering the desire goes together. Let us consider the case according to the single process account. I take it that it proposes two different stories here. One story has it that Hazel did not really forget the second-order belief, but it simply has not been occurrent for a very long time. The second story is that Hazel indeed lost the second-order belief, and know discovers her desire in virtue of self-directed mind-reading. The question now is whether we mix intuitions from these two different stories. We get the intuition that Hazel could easily discover her desires from the first story, in which the single process model predicts that she can easily report her mental state, because the dispositional belief only has to become occurrent. However, we conflate this with Hazel forgetting the belief as in the second story. Here the single process model predicts an alternative method of self-ascription in this case, namely self-directed mind-reading, which might not be as easy (depending on the available evidence).

Second, suppose the intuitions are not the result of mixing two different cases. She actually forgot the second-order belief and she still can easily discover her desire. In this case only the second story is available. Hazel has to know her desire by self-directed mind-reading. The worry here is that I cannot simply propose that mind-reading has no relevant differences to knowing via the single process. If there were no differences then it would be better to only accept the mind-reading route. So what are the differences? One of them can actually be the ease of reporting one's mental state. Self-ascriptions with the single process model have all the resources available that the attitude-forming process uses. Mind-reading at a later state has different resources available because it occurs at a moment that is separate from the attitude formation. The basis for the attitude might not be available as a basis for mind-reading anymore, but it was available as the basis for self-ascription in the single process model. This difference is enough to dispute that single process and mind-reading is equally easy. Moreover, it is enough to claim that there likely is a difference in reliability. To provide a helpful metaphor: We might say that according to the single process model the self-ascription of an attitude is close to the attitude, whereas mind-reading is

[99] This worry was raised by Alex Byrne.

one step further away. The former is related to the attitude formation. The latter has no such direct connection. Hence, self-ascribing by a single-process is expected to be more reliable and easier.

The second case I want to consider is a case of third- or higher-order attitudes. Given the structure of the single process model we can ask whether higher-order attitudes can also be put into the position of the first-order attitude in (SPDM). If they can without restriction, then the model is committed to an infinite number of dispositional beliefs, which is rather implausible. N^{th}-order beliefs would still be generated by the single process that also produces the $N+1^{th}$-order belief. The question at which level the reiteration of the model stops is an empirical one. There is nothing stopping further iterations in principle, even though they would come with additional costs in terms of cognitive resources used. Nevertheless, I lean towards the claim that the single process generates second-order beliefs at most. One way to argue for this is to consider evolutionary advantages of having second-order beliefs quickly accessible (e.g. in coordinating with the group by expressing one's beliefs about one's intentions), whereas there is less of a benefit for third- or higher-order beliefs.[100] With this idea we can provide a just-so story that explains the development of automatic generation of second-order beliefs when one forms first-order attitudes. It simply turned out to be an advantage to have knowledge of one's own mental states. Moreover, the cognitive resources required were plausibly far lower than the benefit of communicating one's mental states.

Hohwy (2013) proposes a further plausible evolutionary advantage of second-order beliefs based on the functioning of our cognitive apparatus: they are important for action representation. Beliefs about one's mental states are not only communicatively relevant, but they are required to plan potential actions and simulate potential outcomes. Without representing one's intentions and one's beliefs, successful planning of action seems impossible – both for groups and individuals. We can compare this benefit of second-order beliefs that Hohwy emphasizes with the need for third-order beliefs. I take it, that representation of one's own second-order belief can be useful, but it is certainly not even

[100] There is some (though weak) empirical support for the claim that second-order beliefs about one's mental states are produced differently than third- or higher-order beliefs about one's mental state. Studies of Jason A. Wheeler Vega (2007) show that subjects have troubles understanding monadic higher-order propositional attitudes (attitudes involving only one individual) above second-order. Given that self-ascriptions are monadic, one explanation for this might be the single process model generating second-order beliefs, but not higher-orders.

close to the importance of representing first-order states when it comes to action planning. To plan my action I need to represent the goal and available means, I do not need to represent what I think about the goal and available means. Hence, we can expect significant benefits of having automatic second-order belief production, but less benefits for third- or higher-order belief generation.

Suppose these remarks are correct. How can we then form third-order beliefs? How can I believe that I believe that I believe that p? The only available answer for the single process model is again to refer to self-directed mind-reading. Higher-order mind-reading seems to be possible in general. I can know what someone else knows about her own mental states. It should be possible to know what I know about my mental states in the same way.

4.8 Conclusion

I argued for a single process model of self-knowledge. First-order attitudes and dispositional, second-order beliefs are generated by a single process, if everything goes right. This gives us an explanation for the apparent asymmetry, reliability, fallibility and transparency. The model also fits with the phenomenology of self-knowledge. Furthermore, with the inclusion of the distinction between dispositional beliefs and occurrent beliefs, the single process model can avoid the standing state problem for transparency accounts.

What are missing here are a further analysis of the single process model for nonintentional, experiential states and a developed discussion of the attitude-forming process. Moreover, the account needs an explanation of the cognitive mechanisms in play, and a developmental story. I will provide an attempt of a cognitive story in chapter 6. However, I remain largely silent on developmental issues. The goal here was to provide a model and show its explanatory power. This explanatory power is the motivation to search for a cognitive and developmental account that fits the model. It might turn out that no account is adequate. Nevertheless, I provided reasons for us to start searching.

5 Extending the Single Process Model

In this chapter I extend the single process model to experiential states such as being in pain, or feeling cold. In doing so I discuss whether the single process model is compatible with the intuition for luminous experiential states. I go over different formulations of the luminosity claim and three different arguments against luminous states. I argue that the single process model is incompatible with proper luminosity and also incompatible with a weaker notion that only proposes that experiential states come with propositional justification that one is in the experiential state. Nevertheless, the model can explain why we have the luminosity intuition in the first place. I finally present the resulting principle of the single process model for experiential states.

5.1 Introduction

In the last chapter I introduced the single process model for propositional and non-propositional attitudes. However, to provide a fully unified transparency account of self-knowledge the model has to be extended for mental states that are not attitudes – nonintentional[101], experiential[102] states such as being in pain, or feeling cold. Beliefs about one's own experiential states are thought to be more likely to be true than beliefs about one's attitudes. Moreover, experiential states are often said to be luminous: one is in a position to know that one is in a certain mental state simply in virtue of being in that state.[103] The major part of this chapter is to find out whether the single process model for experiential states is compatible with a luminosity claim. I first provide a general introduction to luminosity with three common challenges. I then argue that no form of proper luminosity fits with the assumptions of the single process model. However, I show how knowledge of experiential states might nevertheless behave similar to the idea of luminosity. I provide a single process model for experiential states that shows *pseudo luminosity*. Finally, I explain how this model accounts for the comparatively high likelihood of having true beliefs about one's experiential states.

5.2 Luminosity

In chapter 1 I introduced Crispin Wright's (1998; 2015) characterization of self-knowledge in terms of features of avowals, speech acts of authoritative, psychological self-ascriptions. One of these features that Wright observed in our ordinary linguistic practice was *salience*.

[101] I bracket positions that propose that all experiential states are also intentional states, such as developed by Tye (2000).
[102] I am going to use 'experiential' and 'phenomenal' interchangeably.
[103] This feature is also referred to as (some types of) mental states being self-intimating.

Salience captures the idea that whenever one is in a mental state it is absurd for the person to claim to not know whether she is in that mental state. Wright took this to be observable in our everyday practice. If I ask a friend whether she is in pain, it seems indeed odd for her to state 'I don't know.' In the beginning of the last chapter I challenged the luminosity claim for attitudes based on arguments for fallibility and observations of our linguistic practice. I claimed that it seems appropriate to say 'I do not know whether I want x' in some circumstances. However, Wright (1998; 2015) already accepts that salience appears more plausible for phenomenal avowals, and he is not alone in his observation. It is a common assumption in the literature on consciousness that only experiential states, come with awareness of one being in these states. This assumption arguably traces back to ideas as early as in the writings of Aristotle and Buddhist traditions (Strawson, 2015).[104] On the level of belief salience is captured with the term luminosity: one is in a position to know that one is in a certain mental state simply in virtue of being in that state. Luminosity comes in different formulations. For instance, Strawson (2015, pp. 8-9) provides the following attempts:

- All awareness involves awareness of that very awareness.
- All consciousness involves consciousness of that very consciousness.
- All experience involves experience of that very experience.
- All experiencing involves experiencing of that very experiencing.
- All experiencing involves experiencing that very experiencing.

To what degree these formulations differ depends on the interpretations of 'awareness,' 'consciousness,' 'experience' and 'experiencing.' Strawson thinks that some of these formulations sound more plausible and appropriate in our ordinary language. For instance, he points out that the term 'of' can be easily read in a way that suggests a distance between two states, whereas the final formulation avoids this potential issue by dropping the word 'of.' He settles on the last of these using the notion of 'experiencing' and argues that this is also a guide to understanding other definitions using 'awareness.' For my purpose we can treat all these formulations equivalently. Importantly, all these notions assume an implicit, built-in knowledge condition according to Strawson. Hence, he explains when talking about self-intimating states:

[104] Proponents can be found throughout times and traditions: e.g. Śrīharṣa (Das, 2018), Descartes (2008 (1641)), Brentano (1995 (1874)), Husserl (1991 (1907-09)), Goldman (1970), Frankfurt (1988), Weatherson (2004).

> The core meaning of 'to intimate' is to make known and the sense of 'know' in question here is the fundamental sense given which it is correct to say that when it comes to knowing what experience of pain (say) is like, the having is the knowing. This is knowledge in the sense of direct acquaintance (Strawson, 2015, p. 11).

His formulations are as strong as they get. "The having is the knowing" is the adequate slogan for the thesis of luminosity in his view. Whenever one has an experiential mental state, one also knows that one has such a mental state.

Compare these formulations to the prominent definition of the luminosity condition by Williamson:

> For every case α, if in α C obtains, then in α one is in a position to know that C obtains (2000, p. 95).

In this definition 'C' refers to any condition, including conditions such as feeling cold. For my purpose here we can read Williamson's formulation of luminosity as the claim that in every case in which one is in an experiential mental state, one is in a position to know that one is in that state. This is weaker than Strawson's notion, but still claims that everything needed to get knowledge of a phenomenal state is already present merely in virtue of having that state. Nothing is missing. Williamson is deliberately vague about the notion of 'being in a position to know' because he thinks that various ways of spelling it out are compatible with his aims. In the language of a JTBx account of knowledge I will interpret it as the claim that a subject in a phenomenal state is able to form a non-lucky, justified, true belief in their current position. No further inquiry or concept acquisition is needed. Strawson's formulations entail Williamson's, because knowing trivially entails being in a position to know.

There are some reasons to be skeptical of luminosity. We already came across Snowdon's (2012) case against Wright's feature of salience in chapter 1. Snowdon provided the following counterexample:

> For example, when having my eyes tested I am asked which of two lenses results in the greater blurring. It can be very hard to say, and I can aver that I do not know which is blurrier. Now, this is a comparative judgement which relies on memory, but it would not be true to the experience to suppose that the worry is generated by not remembering. The problem is, rather, that it is hard to judge which is blurrier. This seems to be a phenomenal judgement about which one can aver ignorance (Snowdon, 2012, p. 252).

The case supposes that one experiences two different states of blurriness, without being able to discriminate properly between them. If having an experience is knowing it, then one ought to be able to distinguish these two states of blurriness. We cannot, do that sufficiently, so having the experience seems insufficient to fully know it. The same line of argument can be raised against luminosity as being in a position to know.

For the friend of luminosity at least two options are available. First, one can opt for a move similar to Wright (2015) suggesting that only in good cases having the experience is knowing it, where a case is 'good' insofar as nothing interferes with one's ordinary cognitive capacities. The second option is to concede that having an experience does not constitute full knowledge, but merely knowledge of an experiential type. The idea is that having a blurry experience is enough to know that it is a blurry experience, but not enough to precisely tell what kind, or what grade of blurry experience it is. However, this latter option might lead into problems akin to Williamson's anti-luminosity argument, which I will cover shortly.

A second group of potential counterexamples to luminosity comes under the label of inattentional blindness, coined by Mack and Rock (1998).[105] Inattentional blindness characterizes cases in which a person is not consciously aware of a stimulus due to a lack of attention to that stimulus. The most popular case of inattentional blindness is a study involving a person dressed up in a gorilla costume walking through a video. The costumed person is not recognized by the study participants due to their attention being occupied by other tasks (Simons & Chabris, 1999). A possible explanation of these cases is that they involve experiences without awareness. The participants have the visual experience of the gorilla, but they are not aware of that experience. This explanation is not limited to cases involving visual perception. A similar story can be told for injured athletes that do not notice pain until their game is over (cf. Shoemaker (1994)). Again, a possible description of the case accepts that the pain is present, even though the athlete is not aware of it due to a different focus in attention. The difficulty with inattentional blindness cases is that the anti-luminosity reading does not seem to be any better than other description of the cases. A friend of luminosity can simply describe these instances differently. The participant in the

[105] Inattentional blindness is related to change blindness. Both are failures of visual awareness. Change blindness is the failure to notice a change. Inattentional blindness is the failure to notice the existence of an object. For an overview of these two types of blindness and their relation see Jensen et al. (2011).

gorilla case did not have a visual experience, and the athlete did not experience pain. We are only tempted to ascribe these experiences to the participants and the athlete because we expect people in these situations to have the respective experiences. However, the proponent of luminosity can argue that we are simply misascribing the experiences in both cases. It is not that the case that they have an experience they are not aware of. They don't have the experience that we expect them to have.

A third argument against luminosity is provided by Williamson (2000). He imagines a situation in which one feels freezing cold, but slowly warms up until one feels hot later in a series of one millisecond short steps. If feeling cold is luminous, then one ought to be in a position to know at each instance whether one feels cold. For any point luminosity holds that:

1) If in α_i one feels cold, then in α_i one is in a position to know that one feels cold.

To simplify the argument Williamson's story also includes the idea that one constantly concentrates sufficiently hard on one's phenomenal states in the scenario. Hence, one actually knows and is not merely in a position to know. In the imagined case the following is supposed to hold if we accept luminosity:

2) If in α_i one feels cold, then in α_i one knows that one feels cold.

Moreover, Williamson argues that for one to know that one feels cold, one must be within a margin of error. Suppose that one knows that one feels cold at α_i. In this case one must also feel cold at α_{i+1}, otherwise the judgment at α_i would not be reliable enough, given that the case is constructed such that one cannot tell apart α_i from α_{i+1} because the experience at these two points in time are extremely similar. Hence, Williamson gets

3) If in α_i one knows that one feels cold, then in α_{i+1} one feels cold.

As soon as these are established, he can show that for any α_0 in which one feels cold and knows that one feels cold we can get to an arbitrary α_n at which one also feels cold and knows that one feels cold. If knowing that one feels cold at α_0 requires feeling cold at α_1, knowing that one feels cold at α_1 will also require the person to feel cold at α_2 and so on. But the case was described as the person warming up until they feel hot, so it must be wrong that the person feels cold at every α_n. Something must have gone wrong. Williamson

suggests that the idea of luminosity itself is the one to blame. Hence, we should reject luminosity for all mental states, even experiential ones.

I consider two different options for the proponent of luminosity to resist the argument. One option is to reject (3). If feeling cold is luminous, it does not require a margin of error. The self-intimating nature of feeling cold itself guarantees that the judgment that one feels cold is reliable. In particular, this response argues that Williamson's use of a principle of *safety* is misplaced. Williamson understands safety as the condition that if one knows, one could not easily have been wrong in a similar case (Williamson, 2000, p. 147). However, in the case of luminous states safety should not be spelled out in this way. Weatherson (2004) argues that safety ought to be spelled out in terms of the same belief being true in all similar worlds. The crucial difference here is that the condition refers to the belief. If an experience is luminous, meaning that having the experience is having the belief, one cannot be in a similar world with the belief, but without having the experience. The reason for this is that the belief itself depends constitutively on the experience. One cannot have the belief without the experience. So there cannot be any possible worlds with the very same belief without the experience. Hence, Weatherson (2004) argues that (3) is false for luminous states. Instead we get a different safety condition for the belief that one feels cold:

The belief B that one is cold is safe, iff the belief B is true in all similar worlds.

Here the belief B includes the experience itself. Hence, in any possible world in which the belief B exists, the experience also exists. This comes with the drawback that some beliefs are not merely related to, but constituted by experiences – something that requires more, independent motivation than Weatherson provides. Similar attempts of arguing against Williamson's interpretation of the sensitivity condition that leads to a margin of error principle have been provided by Berker (2008) and Vogel (2010). Srinivasan (2015) argues that Williamson's interpretation of the safety principle is correct, and therefore any objections to the anti-luminosity argument on the basis of an interpretation of the sensitivity condition fail.

A different kind of response for the friends of luminosity is to weaken the luminosity claim. I considered a position that conceded that experiential states are not fully, nor always luminous already when discussing possible responses to Snowdon's (2012) argument against the salience of phenomenal states. A related, though slightly different, concession

can be made as a fallback responding to Williamson. DeRose (2002) suggests that one might hold on to a weaker version of luminosity that only functions in cases in which Williamson's safety principle is satisfied. He proposes the following formulation for weak luminosity:

> For every case α, if in α C safely obtains, then in α one is in a position to know that C obtains (DeRose, 2002, p. 576).

Williamson's argument relies on borderline cases, but the weaker alternative simply rules out all borderline cases as relevant for luminosity. All that is required is that experiential states are luminous in clear cases. In the good cases having the experience is enough to know it. Borderline cases just happen to be bad cases in which the experience alone is not enough. The question is how much explanatory power this approach provides. One might worry that the way to identify whether C safely obtains comes back to identifying whether one knows that C obtains. In that case the condition becomes trivial and loses its explanatory purpose (McGlynn, 2014).

A similar fallback, proposed by Berker (2008), concedes that luminosity (respectively the weaker condition called *lustrous* instead of luminous) does not guarantee knowledge, but merely justified belief. This fallback also comes with the price of losing out on the slogan that having the experience is knowing it, but this slogan might be too ambitious anyway.

5.3 A Luminous Single Process Model?

Given these arguments against luminosity, is a single process model including luminosity feasible and how might it look like? I argue that there is no place for the strong conception of luminosity that claims that experiential states necessarily put one in a position to know that one is in such a state. Moreover, even a weaker notion of luminosity seems to be incompatible. I show that the only luminosity that is compatible with the single process model is no actual luminosity at all.

5.3.1 Strong Luminosity

I call the proposal that experiential states necessarily put one in a position to know that one is in such a state 'strong luminosity.' Strong luminosity can be spelled out in at least three ways. First, the experiential state is at the same time the knowledge that one is in this experiential state; or second, the experiential state itself comes necessarily with a ontologically distinct belief that one is in the experiential state that qualifies as knowledge;

or third, the experiential state provides you with all the tools necessary to form a belief about the experiential state that qualifies as knowledge.

The first option requires a metaphysical picture that allows for a single mental state to be two very different things: an experiential state, and the representation of the experiential state. This is the same sort of picture that was discussed in chapter 3 based on Boyle (2011). Boyle proposed that a belief and knowledge of that belief could be a single psychological state. A friend of luminosity can use the same idea to propose that experiential states are also knowledge of themselves, whereas knowledge of the experiential state is taken to be another aspect of the state itself. If I break an ankle I am not only in pain, but I can be *knowingly* in pain by attending to the pain. Being in pain knowingly is not a distinct belief about being in pain, it is merely the state of being in pain under a particular aspect of presentation. If we add the idea that phenomenal states are necessarily attended to we get strong luminosity. Because nothing besides attending to the pain itself is needed for being in pain knowingly and phenomenal states are necessarily attended to, pain is a luminous state. And the same can be said for all experiential states. It is not clear why one should accept this metaphysical picture in general, but I certainly cannot accept it if I want to hold on to the single process model. The single process model requires two distinct mental states: A first-order mental state, and a dispositional belief about the first-order mental state. The metaphysical picture presupposed by the single process model is in conflict with the first option of supporting strong luminosity.

The second option is to allow for distinct states, but argue for a constitutive relation between an experiential state and a belief about that state, such that the experiential state necessarily comes with knowledge of being in that state. Smithies (2012a) describes the idea as a doxastic version of constitutivism that holds the following principle:

> [F]or some mental states M, necessarily, one is in M if and only if one believes that one is in M (Smithies, 2012a, p. 264).

I can only be in pain if I believe that I am, and I can only believe that I am in pain, if I actually am. The principle entails both luminosity and infallibility. Luminosity is the conditional from left to right, and infallibility the conditional from right to left. Hence, one knows that one is in pain whenever one is in pain. However, this version seems false in obvious ways. I presented cases of inattentional blindness and Snowdon's case of the eye test before. These speak against the principle. However, one can formulate weaker principles that have

a better chance of holding up against counterexamples. A better version includes conditions of competence, attention, rationality, and having the required concepts. Smithies formulates it as follows:

> [F]or some mental states M, necessarily, if one is conceptually competent, attentive and rational, then one is in M if an only if one believes that one is in M (Smithies, 2012a, p. 268).

This certainly is a better formulation.[106] It can rule out the cases of inattentive blindness with the 'attentive' clause, and Snowdon's case by claiming that we are not conceptually competent enough to distinguish between two different but very similar visual experiences. However, this is still not an option for luminosity in a single process model.

The third option is to claim that the experiential state itself ensures that one is conceptually competent, attentive, and rational enough to form a non-lucky, justified, true second-order belief that one has the experiential state. Nevertheless, the first-order state does not come necessarily with the second order state. All the ingredients for knowledge of the experiential state are present in virtue of the experiential state; they just need to be used to form the second-order belief. That an experiential state already guarantees that one has the conceptual resources present might be explained by making second-order beliefs referring to the experiential states with demonstratives ('*this* red experience') or by arguing that all that is needed to acquire phenomenal concepts is to experience the respective phenomenal state, a move sometimes employed in discussion of anti-physicalist arguments. Consider Frank Jackson's (1982) knowledge argument, in which Mary learns everything physical about color while living in a black and white room without ever experiencing any colors. When she moves out of her room and sees a red object for the first time she acquires the concept of phenomenal redness. Plausibly, all that is required for her to learn this concept is to experience a red quale. Hence, the experience alone is sufficient to acquire all concepts necessary for knowledge of being in the experiential state.

[106] Sydney Shoemaker has similar qualifications in place. He writes with regard to belief: "[…] if one has an available first-order belief, and has a certain degree of rationality, intelligence, and conceptual capacity (here including having the concept of belief and the concept of oneself), then automatically one has the corresponding second-order belief. (Shoemaker, 1994, p. 288) At times Crispin Wright proposes a constitutivist account that refers to similar C-conditions, such as conceptual resources and sufficient attention. See for instance "Wittgenstein's rule-following considerations and the central project of theoretical linguistic" in Wright (2001).

None of these options look attractive to a proponent of the single process model. My model is set up such that the first-order state and the dispositional second-order belief are not constitutively related. The second-order belief bypasses the first-order state. Moreover, both mental states can be influenced differently by other mental states. That the second-order belief is reliably formed is merely a case of it being sufficiently well correlated with a corresponding first-order state. This is an empiricist account of self-knowledge without any room for constitutive relations between first-order state and second-order belief. As such the constitutivist support for strong luminosity cannot find a place in the single process model. For the same reason there is no room in the model to claim that an experiential state provides all the tools required to come to know that one has that state. However, this does not entail that luminosity has to be given up completely. A weaker form of luminosity is still a live option.

5.3.2 Weak Luminosity

Proponents of weak luminosity (akin to Berker's (2008) notion of lustrous states) accept that the experiential states do not always come with knowledge of these states, nor put us necessarily in a position to know that we are in these states. However, they still argue that there is an important relation between experiential states and the beliefs that we can form about having these states. Experiential states necessarily provide justification for beliefs about the experiential states.

One way to argue for this idea is to identify the experiential state with the justifier for a belief about being in that state. For instance, my being in pain justifies my belief that I am in pain. Smithies' "Simple Theory of Introspection" develops this approach. He argues that "the distinguishing feature of introspective justification is that *its source is identical with its subject matter*" (Smithies, 2012a, p. 261).[107] Whenever you are in an experiential state, you thereby have justification for believing that you are in the state. Nothing else required beside the experience. This by itself does not entail knowledge of all one's experiential states. These experiential states are merely propositional justification that can be used as a basis to form a belief. Only if one forms a belief on the basis of these experiential states, then one's belief is doxastically justified. Hence, having an experiential state is an important step towards being in a position to know that one has that state, but it does not entail that one actually knows. This is an important move to combat counterexamples that threaten

[107] Smithies acknowledges that Neta (2011) develops a similar proposal.

views that take experiential states to entail beliefs about being in the respective states. I showed that one might be unable to form a belief based on one's introspective justification because one lacks certain concepts, or is inattentive, but Smithies argues that in all these cases one still has the introspective justification in place. The experiential state itself guarantees that we have propositional justification for the belief that we are in the experiential state. It simply does not entail that we can thereby come to know.[108]

A related account is developed by Horgan and Kriegel (2007). They propose that some sort of inner awareness is presupposed in our concept of phenomenal experience. This inner awareness is constitutive for phenomenal experience, because it determines the phenomenal character of the experience (Horgan & Kriegel, 2007, p. 134). What it is like for a person to experience something depends on the inner awareness the person has. If my inner awareness is of something sweet, then the phenomenal character my experience has is that of sweetness. Importantly, the inner awareness Horgan and Kriegel discuss is not already a state of belief. Instead they offer the notion of a *proto-belief*, as a less sophisticated doxastic attitude involved (Horgan & Kriegel, 2007, p. 137). However, it is especially easy to form full-fledged belief about phenomenal states because the phenomenal experience already includes the proto-belief. It is so easy that you cannot even form false beliefs for a subset of beliefs about phenomenal experiences.[109] All that is needed is a shift in attention and one can go from proto-belief to belief (Horgan & Kriegel, 2007, p. 137). The view they end up with involves a form of weaker luminosity. Experiential states come with some inner awareness, but one needs to further attend to this awareness to form a belief about experiential states. Again, the experiential states serve as justifiers, but one only gets self-knowledge if one forms beliefs based on the justifiers. And there is no guarantee that one can do that.

[108] Smithies (2016) argues that his account includes that an introspective reason to believe that you're in a mental state is enough for you to be in a position to know that you are in that mental state (Smithies, 2016, p. 359). This is in conflict with a claim in his earlier paper in 2012a in which he explicitly endorses "[…] cases in which one lacks the capacity to use one's introspective justification in forming introspectively justified beliefs" (Smithies, 2012a, p. 266). The 2012a version of his account falls under weak luminosity, whereas the 2016 version appears to have shifted towards a stronger notion of luminosity. I take this difference to come down to different readings of 'being in a position to know.' With a suitable formulation of this phrase even the 2016 version fits into the weaker notion of luminosity. See Smithies (2012b).

[109] They discuss in length for what subset of phenomenal beliefs infallibility is plausible. These details need not concern us here.

Can the single process model adopt weak luminosity in a similar fashion? I do not think so. Horgan, Kriegel, and Smithies start with the assumption that experiential states can be justifiers for beliefs about these states. That is, they presuppose that the belief that one is in mental state M can be based on the mental state M. My proposal of the single process model on the other hand was put forward with the second-order belief bypassing the first-order state. This was motivated by economic advantages and holding on to the transparency idea. A first-order mental state and a second-order belief about that state ought to have the same basis, so that they can be formed by the same procedure. They need to involve the same outward phenomenon to satisfy the conditions for a transparency account of self-knowledge. Hence, forming a belief about one's experiential state based on this experiential state does not fit well with the transparency proposal. Given that my aim is a unified transparency account of self-knowledge I have to reject weak luminosity, because it requires a basing relation of self-beliefs that is incompatible with the single process model.

5.3.3 Pseudo Luminosity

I showed that actual luminosity does not seem to find a place in the single process model. However, this does not necessarily mean that we have to abandon the intuitions that led us to the idea of luminosity in the first place. Wright is correct in his claim that it seems odd to say that I do not know whether I am in pain. I propose that we can accept this intuition at face value without having to accept any sort of luminosity. Instead, the strategy is to explain the intuition as having a different source.

To do this we have to first locate where exactly the intuition comes from. I take it that the intuition is motivated by a difference between knowledge of experiential states to knowledge of non-experiential states. Wright (1998) points out that a luminosity equivalent on the level of language[110] looks different for phenomenal avowals and attitudinal avowals. Only in phenomenal avowals we get the strong intuition that it seems inappropriate to say, for example, that I don't know whether I am in pain. For attitudinal avowals it seems felicitous, at least in some situations, to avow that one does not know whether one has a particular propositional attitude. There is a similar focus on phenomenally conscious states in Strawson (2015) indicated by his final choice of formulating luminosity in terms of 'experiencing.' The common factor behind the luminosity assumption is the intuition that

[110] He calls it *transparency* in Wright (1998) and *salience* in Wright (2015). See chapter 1.

we are better at introspecting phenomenal states than attitudinal states. We are so much better that we expect everyone to be able to tell whether they are in a certain experiential state or not. I want to hold on to the intuition that we are better at introspecting experiential states and explain why this leads to the expectation of people being able to avow them. However, I do not think that this requires luminosity – neither in the strong nor the weaker form. The single process model has the tools to spell out the privilege of experiential self-beliefs in a way that explains our intuitions without the demanding baggage.

My proposal here is an alternative explanation of the intuitions. For this explanation suppose we are reliable in forming second-order beliefs in general. This is already part of the single process model, so at this point one should be happy to take this on board. The strategy now is to look at possible differences between beliefs about experiential states and beliefs about non-experiential states. One proposal would be to argue that one of them is more reliable. Perhaps the single process generating an experiential state and a belief about the experiential state produces a true second-order belief in almost all the cases. In this case it looks as if we have a sort of luminosity. It is not rooted in a constitutive relation between the mental states, but it would be just an empirical fact that one has a justified true belief about one's experiential states. However, this proposal is no good. Even though it involves something similar to luminosity, it cannot make good on the observation that there is something wrong with asserting 'I do not know whether I am in pain.' After all, second-order beliefs in the single process model are dispositional beliefs. If they are, then it should be possible for me to have the dispositional belief that I am in pain, without having the occurrent belief that I am in pain. Given that occurrent belief is a necessary condition for avowing, it seems perfectly possible for me to be in a state in which I cannot avow that I am in pain, or that I am not in pain. But this situation was supposed to be absurd! The intuitions behind luminosity demand an explanation of one being in a position to avow that one has an experiential state. Any proposal that centers on producing dispositional beliefs will not do on its own, and perhaps is not even necessary at all.

My preferred solution is to look at the conditions for dispositional beliefs becoming occurrent. In chapter 4 I provided these four conditions for dispositional beliefs:

(i) S has endorsed the content of p;
(ii) S has stored this content;

(iii) S can recall this content in the right circumstances; and
(iv) the content of p affects S's behavior, reasoning and mental states in the right circumstances

My suggestion now is this: experiential states interact with the triggering conditions for dispositional beliefs to become occurrent in a way that attitudes do not. More precisely, I suggest that the phenomenal experience, the 'what it is like to be' of a state is part of the circumstances that make a belief about that state become occurrent. Consider the following case according to the single process model to illustrate the idea: A single process generates the experiential state 'pain,' and at the same time produces the dispositional second-order belief that I am in pain. The experiential qualities of 'pain' themselves constitute circumstances that make the dispositional second-order belief become occurrent. Pain is an occurrent state, and hence it should not be surprising that it is part of the circumstances that potentially trigger dispositional states to become occurrent. In this case there is one dispositional state which's triggering conditions plausibly include the presence of pain: the belief that I am in pain. So given that I have formed the true belief that I am in pain based on the single process model, I will have an occurrent belief that I am in pain. Call the general form of this claim *experiential occurrence* (EO).

(EO) True second-order beliefs about experiential states that are formed according to the single process model are occurrent beliefs.

It is important to emphasize that the experiences are part of the triggering conditions, but not part of the belief-formation. Hence one *occurrently* believes oneself to be in pain, because one is in pain. But that does not mean that one *believes* oneself to be in pain, because one is in pain. (EO) is a claim about the activity of a belief, not the basis of a belief. Hence, (EO) is still compatible with the bypass idea in the single process model. (EO) is limited to experiential states, because only the phenomenal qualities enter into the circumstances that can make a belief occurrent. What it is like to be in pain is part of the situation of the subject in a way that propositional attitudes are not.[111]

(EO) is the key to explaining luminosity intuitions without luminosity. If we are reliable, and all our second-order beliefs about experiential states are occurrent, then we have a natural explanation of why we generally are in a position to avow whether we are in pain. Usually,

[111] This still assumes that general cognitive phenomenology is false, an assumption I already pointed out in chapter 1.

one forms the experiential state 'pain,' and the corresponding *occurrent* second-order belief that one is in pain by a single process. In this case one is able to avow that one is in pain. Is this enough to explain that there is something wrong with the assertion 'I do not know whether I am in pain'? Partially, I think. The proposal states that if one is in pain, then one is in an excellent position to avow that one is in pain. According to the single process model avowals about one's mental states are formed in two steps: First one generates a dispositional belief that one is in a mental state based on the same process that generates a first-order mental state. Then this dispositional belief becomes occurrent in the right circumstances and therefore avowable. For experiential states this second step comes for free, because the state itself provides the circumstances. So a particular way in which you could be unable to avow that you are in pain is ruled out in virtue of being in pain. It cannot be the case that you have the dispositional belief to be in pain, but that belief fails to become occurrent.[112] So there is a way in which you might not be able to avow whether you believe something that is ruled out for the case of pain. However, there are two other possibilities in the single process model: one might form a false second-order belief or no second-order belief at all. If the single process model is right for the normal case the latter should be very rare. Hence, in everyday talk we presuppose that one has formed a second-order belief.

The former case on the other hand cannot be dismissed so easily. The single process model allows for fallibility, and I have not provided any reason to deny this feature for forming beliefs about experiential states. However, I do not need to. Instead, we can employ the same idea of the experiential qualities being related to the triggering conditions of the dispositional second-order belief. Suppose I form the experiential state 'being cold' and the dispositional second-order belief that 'I am nervous' by a single process.[113] Clearly the second-order belief is false. However, if the same mental state forming process generates coldness and the belief that one is nervous then one's cognitive makeup is likely such that the experiential state of coldness will be part of the triggering conditions that make the belief that one feels nervous become occurrent. Given that something (e.g. other mental states) interfered with the production of the correct second-order belief, it will likely also interfere with the triggering conditions. If this is right, then one almost always ends up with

[112] Compare to the therapy case discussed under fallibility in chapter 4.
[113] Suppose these are two different existing phenomenal states for the sake of the argument. Nothing hangs on the particular example chosen.

occurrent beliefs about one's experiential states, regardless of whether they are true or false. And if one almost always ends up with occurrent beliefs, one is almost always in a position to avow what experiential states one is in. This explains why it is so odd to hear someone say 'I do not know whether I am in pain.' There has to be something very unusual going on for this to be the case. Generally one can avow one's experiential states in virtue of having an appropriately occurrent belief. And one can avow that one does not have an experiential state in virtue of not having an occurrent belief that one has the state. So if someone says that they cannot make either of those avowals something has to be wrong.

This explains the intuitions behind the luminosity idea without accepting any form of luminosity. One can have an experiential state without knowing that one has the state. One can in principle even have an experiential state without being justified that one has the state. There might not be a sufficiently high correlation between experiential states of particular type and beliefs about being in states of that type. However, beliefs about one's experiential states formed according to the single process model are almost always occurrent beliefs. And this almost always occurrent nature explains our intuitions behind luminosity. This also shows that beliefs about one's experiential states are not different in kind to beliefs about one's attitudes. They only differ in degree insofar as the triggering conditions for the former are usually met in virtue of having experiential states. Their production follows the same structure.

5.4 The Single Process Model for Experiential States

I showed how we can adequately explain the luminosity intuitions without buying into the luminosity claim. Doing so helps to avoid the problems raised against luminosity in the beginning of the chapter. I am now in a position to formulate the general single process model for experiential states:

(SPE) Normally, a single mental state forming token process P produces both my first-order experiential state Q and my occurrent belief B that I have experiential state Q*. The production of B can be influenced by mental states M, which accounts for the possible difference of Q* and Q. Q* is identical to Q, if everything goes right. Neither Q, Q*, nor B are part of the attitude-forming process.

The qualifying 'normally' acknowledges that in rare cases the belief might not be occurrent, but dispositional. (SPE) is structurally the same as (SPDM) proposed in chapter 4. The important difference is that beliefs about one's experiential states are normally occurrent, whereas beliefs about one's attitudes are normally dispositional. This is a result of the experiential qualities being a part of the triggering conditions for the beliefs to become occurrent.

I explained the single process model as having two stages: a stage of forming the mental states, and a stage of the formed belief becoming occurrent. The idea that these stages can fall into one if the first stage generates the conditions under which the second-order belief becomes occurrent explains the luminosity intuition. However, it can also explain another intuition: ordinarily we think that beliefs about experiential states are more reliable than beliefs about attitudes. Wright (1998), for instance, describes sincere, competent, phenomenal avowals as guaranteeing truth, whereas sincere, competent attitudinal avowals could be erroneous. I am not on board with this strong claim, especially when considering Snowdon's (2012) convincing counterexample presented in chapter 1. However, the single process model for experiential states can accommodate the basic idea that avowals of experiential states are less likely to be based on a particular error. I cannot show that the belief-forming processes involved are more reliable in the experiential case in general,[114] but I can show that one way to end up with false avowals is ruled out. This again builds on the beliefs about experiential states being occurrent. When my belief that I am in pain is occurrent, I usually can make an avowal based on that belief. In virtue of this occurrent belief it is ruled out that I avow on any other basis. If I have an occurrent belief that I can use for avowing, I thereby have no need to confabulate an avowal about my experiential state. In chapter 4 I showed that the single process model can explain at least some confabulation cases as instances (e.g. split-brain cases) in which one has a dispositional belief, but these cannot become occurrent. One then confabulates to come up with a false avowal, if one is asked to report on one's mental states. For avowals about experiential states this is at best a very unlikely error possibility. Normally, one has an occurrent belief in virtue of the experiential state, and therefore can avow that one has the experiential state. This does not guarantee the truth of the avowal. It merely shows that one usually avoids a certain sort of mistake. One avoids any error that is caused by the

[114] This is a job for psychologists and neuroscientists.

dispositional state not becoming occurrent. Hence, the single process model explains the intuition that phenomenal avowals are more reliable than attitudinal avowals. We simply do not make a particular kind of mistake when avowing experiential states.

5.5 Conclusion

In this chapter I extended the single process model to experiential states. The basic structure of the model stays the same, while the intuitive differences between knowledge of experiential states and knowledge of non-experiential states are explained by differences in making second-order beliefs occurrent. The result is a model that fits decently with what Boyle calls the 'Uniformity Assumption' (Boyle, 2009, p. 141): the idea that an account of self-knowledge should explain every form of self-knowledge, and not only a limited number of mental state types. I can even uphold a very strict sense of uniformity "[…] explaining all cases of 'first-person authority' in the same basic way' (Boyle, 2009, p. 141). This was one of the main goals for my account. Most importantly, it was a goal that most transparency accounts cannot reach. As an objection it has been raised against various transparency accounts under the guise of 'the problem of scope' (c.f. Gallois (1996), Finkelstein (2003), Gertler (2011a), Cassam (2014)). The single process model can effectively deal with this problem and explain knowledge of attitudes and experiential states.

It is unclear how much force the uniformity assumption has. Some, such as Boyle (2009), Moran (2001), and Carruthers (2011) happily drop the uniformity to some degree from their accounts. Others, such as Bar-On (2004), Byrne (2018), Finkelstein (2003), and Nichols and Stich (2003) hold on to it. Nevertheless, uniformity is evidently an advantage in explanatory power. If one needs to choose between two accounts that both can account for the apparent asymmetry, reliability, and fallibility, one criterion for picking one over the other is whether one theory can explain knowledge of mental state types that the other cannot. I propose that the single process model is applicable to all types of mental state. As such, it has at least one advantage over its alternatives.

6 Self-Knowledge in a Predictive Processing Framework

In this chapter I propose a cognitive story for the single process model based on a framework of predictive processing. Predictive processing understands the brain as a prediction-action machine that tries to minimize error in its predictions about the world. I provide a predictive processing account for self-knowledge starting from remarks on introspection made by Hohwy (2013). I develop Hohwy's picture into a general model for knowledge of one's mental states, discussing how predictions about oneself can be used to capture self-knowledge. I further explore empirical predictions, and thereby argue that the model provides a good explanation for failure of self-knowledge in cases involving motor aftereffects, such as the broken escalator phenomenon. I conclude that the proposed account is incomplete, but provides a valuable first step to connect research on predictive processing with the single process model and the epistemology of self-knowledge in general.

6.1 Introduction

In this chapter I provide a cognitive story that fits with the single process model. The proposal is based on the idea of predictive processing. Predictive processing understands the brain as a prediction-action machine that tries to minimize error in its predictions about the world. For this view to evolve into a complete account of human cognition we ought to provide an idea how it can account for self-knowledge – knowledge of one's own mental states. I provide an attempt for such an account starting from remarks on introspection made by Hohwy (2013). I begin with an overview of the predictive processing framework. Part 3 explains Hohwy's attempt to embed introspection in the framework. Part 4 develops his picture into a general model for knowledge of one's mental states. I thereby also provide some remarks on the relation of propositional attitudes to the predictive processing picture. Part 5 discusses empirical predictions and provides one case which fits with my proposed account. In part 6, I summarize the account and relate it to questions in the wider philosophical discourse on self-knowledge and the single process model. I conclude that the account is promising, but incomplete.

6.2 Predictive Processing

The predictive processing framework provides a novel account of at least perception and action, but perhaps the workings of the brain in general. It promises understanding of the brain as a prediction-action machine that constantly predicts sensory input and aims to minimize error in its predictions, thereby reinventing Helmholtz's (1860 (1962)) idea that the function of the brain is best summarized by the slogan of error correction. Importantly,

prediction does not indicate a person predicting something. Rather, the notion of prediction in play is a subpersonal, automatic, probabilistic guessing as part of neural processes. Prediction in this sense is something that brains do and which enables embodied, environmentally situated agents to carry out various tasks (Clark, 2016, p. 2).

The initial idea can be explained as the brain trying to guess the causes behind sensory inputs. States of affairs in the world have an effect on the brain via our senses. The difficult task for the brain is then to figure out what these states of affairs are; based on the effects they have on the senses. It is easy to see that the task is difficult, because a single effect can have numerous different causes. A tree and a picture of a tree might have the same effect on our senses, but are clearly different things. Moreover, a single state of affairs in the world can have various effects, simply because we are related to the state differently. A tree looks different from far away, for instance. Predictive processing explains how our brains solve this problem. The idea is that the brain provides hypotheses about the world. It uses an internal model based on past encounters to predict what is out there. Moreover, the prediction about the world is then tested by predicting what the brain's next sensory inputs will be. Based on whether this prediction is correct, the model is updated.

Say I hear a sound in the middle of a weekend night. Based on previous, similar instances my brain takes the most probable cause of this noise to be my flatmate coming home and closing the front door. In the past when this happened shortly after she would go up the squeaking stairs, hence my brain predicts this squeaking sound to come soon after. And sure enough, a minute later the sound hits my ears. The prediction gets confirmed and its assigned probability increases. However, suppose the squeaking sounds had not followed. In this case the prediction would not fit and the probability of this being my flatmate would decrease. An alternative, more probable hypothesis would be put forward. In this fashion probabilities of hypotheses can be updated based on the error of the predictions. The less error in a prediction, the more the hypothesis gets confirmed. Based on this probability updating the predictive brain can learn by itself without going beyond the perspective of the skull-bound brain (Eliasmith, 2005). The brain uses its own 'bootstrapping' mechanism (Hohwy, 2013, p. 16).

The revolutionary story of this framework is its top-down approach. The brain does not simply represent whatever input it gets, but rather it builds an internal model of the world, predicts the sensory input and then modifies its own model based on the extent the predictions were incorrect. This top-down approach can provide explanations that bottom-up accounts cannot. Binocular rivalry for instance – the case in which each individual eye provides vastly different sensory input and the brain picks one over the other. The top-down approach can make sense of this case as the most probable hypothesis here is that there are two things out there, rather than one thing that looks so different from one eye to the other, so the brain represents only one at a time and neglects the sensory input from the other eye (Hohwy, et al., 2008; Hohwy, 2013). A recent fMRI study supports this explanation (Weilnhammer, et al., 2017). Further evidence for the predictive processing model can be found in its explanation of some mental illnesses as discussed by Clark (2016) and van Schalkwyk et al (2017), or the account of the functioning of the retina by Hosoya et al (2005). Friston (2005) also shows that the predictive processing story is compatible with a range of anatomical facts.

In addition to switching to a top-down generative model, the predictive processing framework employs the same move in frequent iteration. A vast number of top-down predictions function together. Predictive processing models are for the most part hierarchical. Different layers of smaller internal modelling processes are related downwards, upwards and sidewards, whereas usually the higher layers are more general and relate to a longer timeframe. Hence, predictive processing uses a complex network of structured models related by predictions and error-signals (Rao & Ballard, 1999; Lee & Mumford, 2003; Friston, 2008). In case of perception these are tracking the external causes based on unsupervised learning (Kawaro, et al., 1993; Hinton & Zemel, 1994; Hinton, et al., 1995).

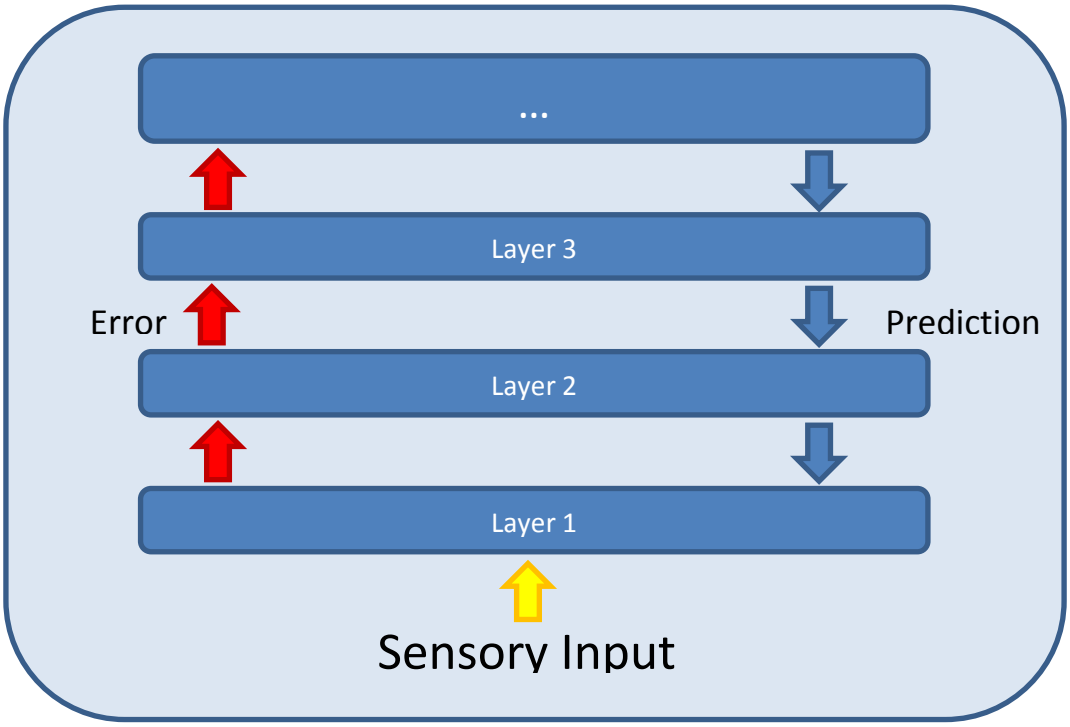

Figure 6

This picture of hierarchical top-down and bottom-up interaction of modeling, predicting and receiving error-signal has been enhanced recently by including action in the same framework (Brown, et al., 2011; Hohwy, 2013; Clark, 2013; 2016; Friston, 2010). The main idea is that predictions can be true in two vastly different ways: either by getting the world right, or by changing the world to fit the prediction. The former one is perception, the latter one action. I can impact the world in a way that reduces my prediction error. Incorporating action is a big step towards predictive processing as a way to understand cognition in a unified way. It would be another step in this direction if introspection can be understood in the same framework. The rest of this chapter aims to explore one way of doing this based on Hohwy (2013).

6.3 Double Bookkeeping

Hohwy (2013) proposes a way to integrate an account of introspection into the predictive processing framework. The idea is aimed at capturing self-knowledge of experiential states and therefore rather limited. However, one can adapt the general structure as a basis to design a general framework for knowledge of one's own mental states, including propositional attitudes. To start let us sketch the model for experiential states. The motivation for the model is the phenomenon of being surprised by experiential events. A pain might be sharper than expected, or a color experience may differ from what the brain

predicted. To capture these phenomena Hohwy proposes a second hierarchical structure that works parallel to the predictive processing story on perception (2013, pp. 245-250). The brains model of the world includes a model of experiences that creates experiential expectations. These expectations are then met with prediction error, which ultimately stems from phenomenal experience. Just as the model for perception, the model for experiences is built hierarchically, where every layer makes prediction regarding the next lower layer and receives prediction errors if these predictions are not met. And just as in the perception case these layers differ in abstractness and temporality. So whereas layer 1 might be about immediate change in visual experiences when I turn my head, layer 3 might be about the gradual changes in experience due to natural lighting conditions (for instance gradual change in lighting in the afternoon). Hohwy is rather vague on the exact mechanisms. However, he gives us a good analogy: double bookkeeping.[115]

Experience is thought of as a byproduct that keeps shadowing perception. Shadowing here is meant to indicate two points. First, that usually perception comes with experience, so perceptual predictions plausibly also come with experiential predictions; and second, that this might go unnoticed most of the time. Both 'books' are in the end based on sensory input. The predictions and error signal of the perception side generate the input that is supposed to match the prediction on the experiential side. We should understand that perception generally goes together with experience as the predictive process of perception being connected to the predictive process of experience. Layers of perception predictions are connected to layers of experience predictions. This is treating the "[…] deliverances of perceptual inference as causes of the input to a model" (Hohwy, 2013, p. 246).

[115] Hohwy (2013, p. 246) considers this first as an objection, but I take it to be a useful metaphor to understand the generation of self-knowledge in his framework.

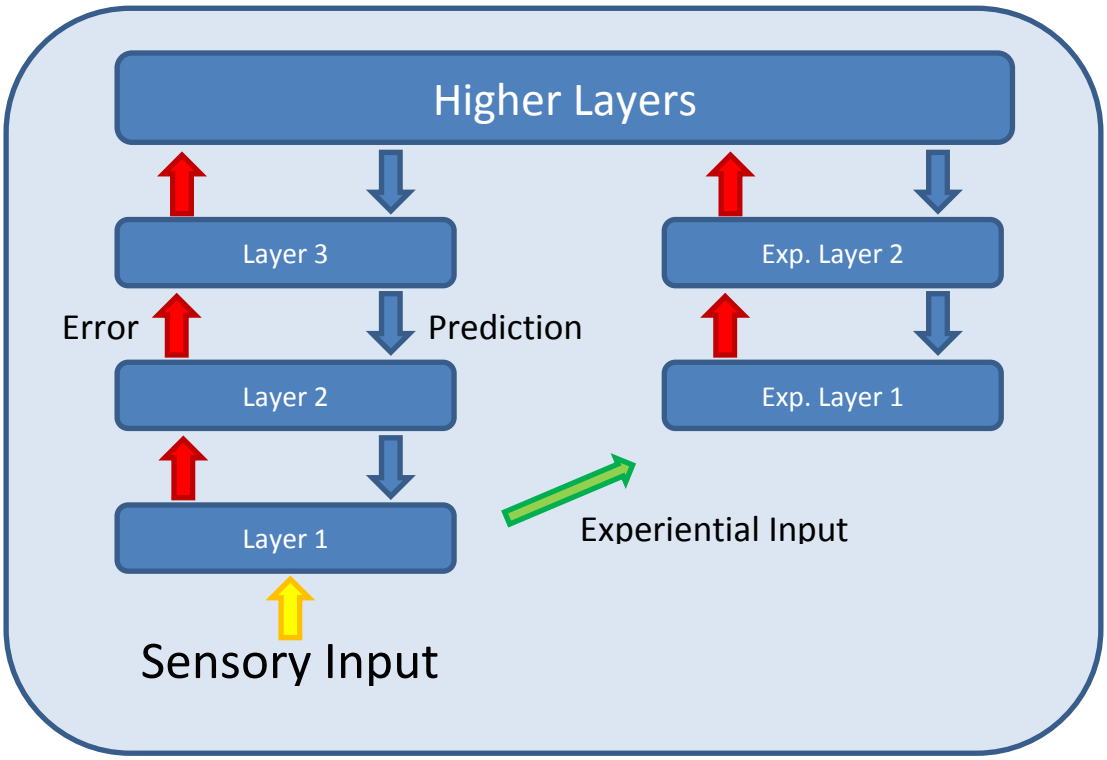

Figure 7

Predictions of experiences are not only based on other layers representing experiential states. They are also connected to predictions about sensory input, and hence predictions about the world. This seems obvious when we consider the role that states of affairs in the world play in prompting experiential states. When I predict a football hitting me on my head shortly I will also predict an unpleasant experience.

Given these connections between predictions about the world and predictions about experiential states, we can explain why experiential prediction errors seem rare. Most of the time the experiential prediction errors go unnoticed, because the error is located on the perception side of processing. We are only aware that something is wrong on the experiential side of things when the prediction errors cannot be explained by inferences on environmental causes (Hohwy, 2013, pp. 247-248). This happens when our expectations of the sensory inputs are correct, but our predictions of the phenomenal experiences are not, for instance in case one is surprised by a very bright light that one turned on. The brain predicts the incoming bright light, but the experiences surpass the experiential predictions. In this case any change in the perceptual prediction will create a worse fit to the sensory input, so this change is off the table. Moreover, any action that changes the input will also create a further mismatch between the sensory input and the prediction of the sensory

input. So this is no option either. The appropriate adaption for the brain is to change the model of the experience that is connected to this particular sensory prediction.

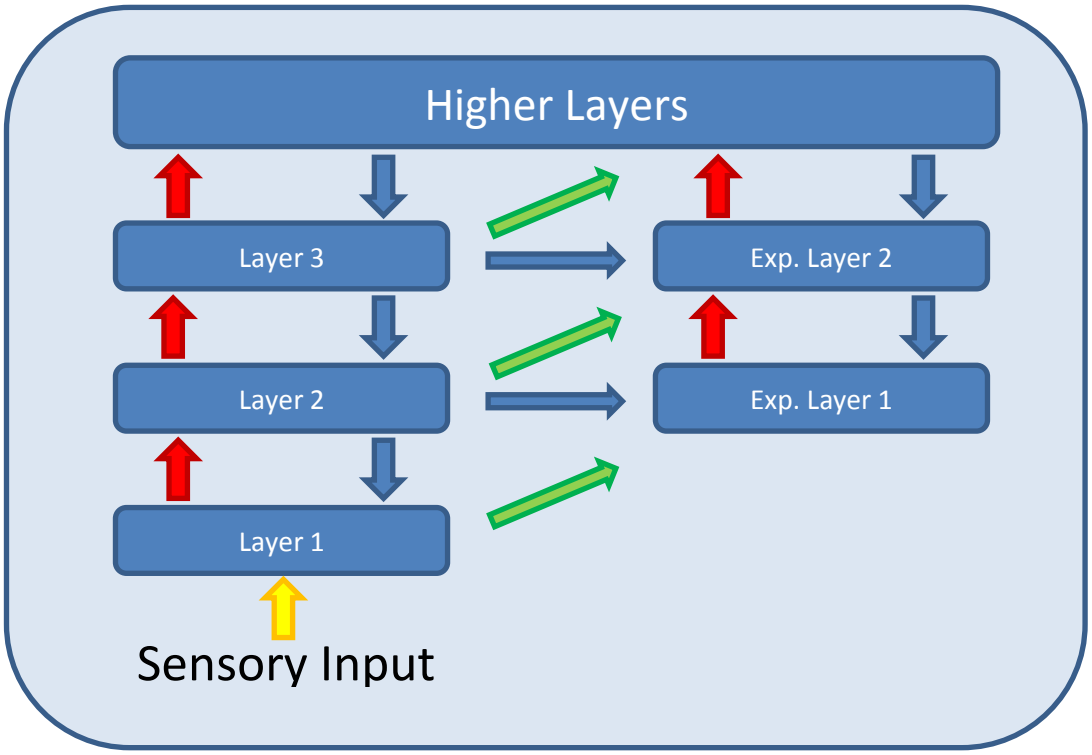

Figure 8

In terms of figure 8, we have a prediction of sensory input in Layer 2, based on Layer 3. We also have an experiential prediction on Exp. Layer 1 based on Layer 2 and Exp. Layer 2. The Layer 1 prediction fits the incoming sensory input, so no relevant error is forwarded to Layer 2. However, the phenomenal experience caused by Layer 1 does not match the prediction in Exp. Layer 1. We have a prediction error at Exp. Layer 1. This error is forwarded to Exp. Layer 2 and (potentially) up to our general model of the world and ourself. The result is an adjustment of the predictions of one's experiences. Ideally, next time one turns on the same light the expected phenomenal experience fits the actual brightness one experiences.

6.4 From Experience to Mental States

Hohwy (2013) introduces introspection with regard to perceptual experiences. However, because active predictive processing combines perception and action he tends to generalize at various points. He writes:

> There is a more fundamental reason to believe that creatures like us introspect. This relates to active inference and our ability to control the environment through

> action. In particular, any agent who represents its own actions must in some sense introspect. Representation of action or control of the environment is a subtle point and may only be a faculty of higher organisms; representing one's own action is not necessary for simple reflexes or homoeostasis—and yet it becomes imperative for creatures like us who engage in planning and entertaining fictive outcomes (Hohwy, 2013, p. 247).

This seems obviously correct, but not accounted for. We can introspect all kinds of mental states and they relate very differently to action. Most importantly we can know what we desire and believe, and we can base our action on these states. Currently the model only features predictions of sensory input (which captures the state of the world) and experiential input (which captures some states of the mind). We need to generalize the double bookkeeping picture to mental states in general to explain self-knowledge completely.

The first issue here is the question how to translate different mental states into a predictive processing picture. The straightforward case is the notion of 'belief.' Predictions share some of the relevant features of belief. Most importantly, predictions aim at truth. However, predictions can be true by different means. If you predict that a glass that you see falling towards the floor will shatter you are correct because your predictions fit the world. It is just a matter of the properties of the glass, the floor and the laws of physics that it shatters. However, you can also predict that the glass standing safely on the table will shatter in the next minute and be correct in virtue of grabbing the glass and throwing it against a wall. In this case the prediction is not true because it fits the world, but because you made the world fit to the prediction. The former prediction shows a mind-to-world direction of fit, whereas the latter shows a world-to-mind direction of fit.[116] It is central to the predictive processing picture that predictions can be true in virtue of either direction of fit. Mental states with either direction of fit can be identified with predictions. Hohwy uses this approach to describe desires:

> What drives action is prediction error minimization and the hypothesis that induces the prediction error is a hypothesis about what the agent expects to perceive rather than what the agent wants to do. If this idea is expanded to standard examples of desires, then desiring a muffin is having an expectation of a certain flow of sensory

[116] Anscombe (1957) provided a fantastic analogy for this difference. It is akin to the difference of consulting a shopping list to select which items to purchase (the list determines the contents of the shopping basket) and listing some actually purchased items (the contents of the shopping basket determine the list).

> input that centrally involves eating a muffin. This means the concept of desire becomes very broad: any hypothesis associated with a prediction error can induce a want or an intention or a desire, simply because such prediction error in principle can be quenched by action.
> What makes the desire for a muffin a desire and not a belief is just its direction of fit. Both are expectations concerning sensory input, and the "motivator" is the same in both cases […] (Hohwy, 2013, p. 89).

There are two different ideas here. First, that a hypothesis together with a prediction error can induce a desire. And second, that a desire is an expectation concerning sensory input with a certain direction of fit. These are two very different things mentioned back to back, but they can go together in the hierarchical structure. Consider a situation in which the brain predicts sensory input from water on the tongue on Layer 1 based on Layer 2. This prediction comes with a mind-to-world direction of fit. Initially, the sensory input does not line up with this expectation, so a prediction error is fed forward. Now the brain has various options to react. A plausible reaction is a new prediction with the same content, but different direction of fit together with an action (grabbing a glass and drinking water from it). Now the sensory input lines up with the prediction. In this case a hypothesis associated with a prediction error induced a desire. The desire is also the new hypothesis, but only as a prediction with different direction-of-fit. The first hypothesis was a belief, and the second, replacing the first, a desire. The desire was fulfilled in virtue of an action based on the desire and background beliefs (other predictions about the environment).

This picture still has challenges ahead. Direction of fit might be good enough to distinguish between belief and desire, but it is not enough to distinguish between different attitudes with the same direction of fit. Velleman (1992) argues that direction of fit is not enough to define beliefs, because "Hypothesizing that p, assuming that p, hoping that p, and the like are all attitudes in which p is regarded, not as a representation of what is to be brought about, but rather as a representation of what is" (p. 12). If this is correct – and it seems correct to me – we need more than just direction of fit.

However, there is a more pressing problem in front of us. A difficulty combining Hohwy's picture with traditional ideas on introspection is that there are different notions of 'belief' in play. Hohwy talks of beliefs and desires, but these are not quite the same things the philosophers of self-knowledge usually talk about. A belief in the predictive processing picture is not necessarily a propositional attitude. One could perhaps make the case that predictions on a high level are structured propositionally, but certainly on lower levels they

need not be propositional. Rather, they are only partial representations that require the other layers to model a state of the world together. They are partial representations insofar as they make predictions and report error signals that are on their own insufficient to represent a state of the world. Call the non-propositional state 'belief$_{pred}$' and the propositional state 'belief$_{prop}$.' 'Belief' without qualifier includes both.

How exactly belief$_{prop}$ relates to belief$_{pred}$ is unclear. Dewhurst (2017) argues that these two conceptions are actually incompatible. The main reason for this is that the folk notion of belief$_{prop}$ is a concept that does not come in degrees, whereas beliefs$_{pred}$ are inherently probabilistic. A belief$_{pred}$ reflects the probability of a certain state of affairs. Dewhurst argues that we need to either reconsider the relation between folk psychology and the brain, or need to revise our folk notion of belief to be probabilistic. He chooses the former while attributing the latter option to Pettigrew (2015). Dewhurst's approach aims to show that we should not use folk notions of propositional attitudes at all when trying to understand cognition on a scientific level that predictive processing does, but instead we should understand the folk notion as a broader behavior interpretational tool. He argues that this is a good way to hold on to the explanatory functions that folk psychological attitudes have in our everyday life. After all, we use folk psychology to explain and predict behavior, and to form narratives for ourselves and others (Dewhurst, 2017, pp. 6-8). Moreover, for my purpose of connecting the predictive processing framework to philosophical accounts of self-knowledge it seems at least instrumentally useful to hold on to propositional attitude talk to some degree.

Fortunately, my proposal does not require a straightforward realist conception of propositional attitudes, even though it would be compatible with that. All I need to provide is an idea how our folk notion of propositional attitudes that is at work in philosophical accounts of self-knowledge relates to predictions in the predictive processing framework. I will opt for an undemanding thesis for the relation of propositional attitudes to prediction and take on a fictionalist stance. The fictionalist stance towards propositional attitudes accepts that even though the ontological status of propositional attitudes might be dubious, it is a useful concept to take on board. This is similar to treating mathematical objects as real, even though it is not fully clear what their ontological status actually is. That predictive processing leads to this fictionalist view of representation in general has been recently argued by Downey (2017). Moreover, there is good reason to opt for the

fictionalist stance. In the predictive processing picture the only causally effective states are predictive states.[117] Propositional attitudes have no place in the causal explanation of predictions and actions. What attributions of propositional attitudes are mostly used for is to explain behavior on a reasonably sized scale in our ordinary linguistic practice. We tend to explain behavior in broad strokes: We explain why someone gets a drink, rather than why someone moves his hand one centimeter after the other. The behavior we usually want to explain is in itself a complex action built up from smaller, more basic actions. This sort of behavior is based on a multitude of predictive states (with both directions of fit), but explaining it with reference to all these different states would be rather inefficient (if not impossible) in our daily communication. Hence, we talk about propositional states that capture these various predictive states working together. The fictionalist stance claims that this is all the propositional states are: tools that let us talk about mental states more easily, even though less accurately. For my purpose it makes no difference whether propositional attitudes are states on ontologically safe ground, or whether they are states that we invented to talk as if they supervene on predictive states. Hence, I can concede to Dewhurst (2017) that our propositional attitude talk has no place on the explanatory level that cognitive scientists aim at and take his view on board. Moreover, with a commitment to fictionalism I can employ a notion of propositional attitudes that is very close to Dewhurst's own proposal. Propositional attitudes still play a role in explaining behavior in our everyday life. Their ontological status is not secure, but that does not make them any less relevant in our folk explanations. Moreover, as long as any account of self-knowledge wants to capture folk intuitions about the asymmetry of the attribution of folk psychological attitudes, we ought to explain how self-ascriptions of these folk attitudes relate to the cognitive processes going on in one's brain. This is still possible if we accept a fictionalist stance on propositional attitudes.

The question of how predictions relate to folk psychological attitudes is still not fully answered. How are they exactly connected? For instance, what $\text{beliefs}_{\text{pred}}$ are relevant for ascribing a $\text{belief}_{\text{prop}}$? I doubt that we are in a position to pick out the exact $\text{beliefs}_{\text{pred}}$ in question. However, we can observe what determines our use of $\text{belief}_{\text{prop}}$ in ordinary language and reasoning, and connect this observation to the $\text{beliefs}_{\text{pred}}$. Following

[117] It is a further topic of current debate whether these predictive states themselves count as representational. For instance, Downey (2017) denies this, whereas Gladziejewski (2016), and Wiese (2017) affirm it.

Dewhurst (2017) we can identify the use of folk attitudes in predicting and explaining behavior and in our corresponding talk. Moreover, he points out that we find folk attitudes in our construction of narratives that further help to predict and explain behavior (Cf. Bruner (1990), Hutto (2008)). The important point here is that our use of propositional attitudes in folk psychology seems to be primarily about the identification, ascription, and predictive use of behavioral dispositions. Therefore Dewhurst rightly claims that "propositional attitudes were never meant to refer to fine-grained mental states, but are instead intended to pick out and predict coarse-grained behavioural patterns and dispositions" (Dewhurst, 2017, p. 10). If we now look at a description of our cognitive going-ons in the predictive processing framework we can pick out what is primarily relevant for behavioral patterns and dispositions: sets of interconnected first-order predictions on various levels. The predictions and their relations to each other taken together are responsible for behavioral patterns and dispositions. What folk ascriptions of propositional attitudes are trying to get at is the predictions and their structure that lead to actions. And in the same vein, what self-ascriptions of propositional attitudes are trying to capture is one's own structure of predictions that lead to actions.[118] An ascription of a folk psychological attitude will not allow one to identify the exact predictions in place. However, it will allow one to pick out well enough what a bunch of predictions together do. 'Well enough' here simply means that an ascription of a folk psychological attitude captures a set of predictions at work sufficiently such that it allows us to rely on the folk atittudes for behavioral predictions, explanations, and the building of narratives in our everyday life.

We can now use the idea that propositional attitudes are connected to multiple representational layers in the predictive processing framework to get a better grasp on the problem of different propositional attitudes. Latching onto an idea of Dennett (2013) we can try to explain more complex attitudes and social features in terms of predictive processing. Dennett proposed to understand cuteness in virtue of expected expectations.

> When we expect to see a baby in the crib, we also expect to "find it cute" – that is, we expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as

[118] A close connection to the prediction of one's own action that fits into this picture will be discussed later.

> confirmation that, indeed, the thing in the world we are interacting with has the properties we expected it to have (Dennett, 2013, p. 30).

Dennett aims this idea towards explaining features of things in the world that are not, or at least not obviously, translated into sensory input. Finding something cute is not to be identified with any individual prediction, but rather with a set of predictions. These include both predictions of sense input and experiential predictions. I want to go a step further and use the same idea to get a grasp on propositional attitudes in general.

The central idea is that different propositional attitudes can be identified with different predictive profiles that lead to different behavioral patterns and dispositions. If there is a straightforward predictive processing explanation to finding something cute, as Dennett proposes, then we can adapt the story for finding something adorable, fearing something, hoping for something, etc. Start with two basic prediction modes – predictions distinguished solely by their direction of fit. These predictions have content, but not propositional content. The two modes are supplemented by adjustments on precision-weightings, differences in margin of error for predictions. We can then understand all other attitudes in terms of these two modes and precision-weighting. For instance, we can understand $\text{belief}_{\text{prop}}$ as if it were a mind-to-world directed state that captures a pattern or profile of mind-to-world directed predictions.[119] Moreover, $\text{beliefs}_{\text{prop}}$ involve predictions that are quite sensitive to error signal, compared to other attitudes. Imagining that p, for instance, also captures a pattern of mind-to-world predictions, but these predictions are not, or at most barely sensitive to error signals originating from sensory input on low levels of representation. In other words: Imagining that p is not threatened by a mismatch with the outside world, whereas $\text{belief}_{\text{prop}}$ is. Sensory input can prompt revision of $\text{beliefs}_{\text{prop}}$, but not of imaginings. A $\text{belief}_{\text{prop}}$ can also involve counterfactual predictions, as have been proposed in the predictive processing models by Friston et al. (2012), Seth (2014) and Pezzulo et al. (2015). The idea here is that counterfactual predictions capture how the sensory input (and more general the predictive profile) would change, were we to interact with the world in a possible way. Moreover, this can be supplemented by adding further conditions that relate predictions that are part of what we take to be a $\text{belief}_{\text{prop}}$ to other types of mental states in a functionalist fashion.

[119] These formulations ought to be read with the fictionalist stance in mind.

Take another example. Desires$_{prop}$ can be understood as capturing a group of world-to-mind directed predictions. And desires can be differentiated from intentions by their responses to error signal, that is, by precision weighting. Similar to the belief$_{prop}$ versus imagining distinction, having an intention involves a higher sensitivity to error signal. I can desire$_{prop}$ something while not doing anything about it being the case. A desire can be unfulfilled and still not prompt an action. That is, there can be significant mismatches on world-to-mind predictions that do not prompt a change on any representational layer, nor cause an action. On the other hand, if I intend something then prediction mismatches are reduced by acting. In this manner a desire$_{prop}$ shows different behavioral dispositions to an intention$_{prop}$.

A final instance is fear that p. This would be understood roughly as capturing a group of world-to-mind directed predictions aiming to prevent p from occurring and, plausibly, some mind-to-world directed predictions about p being dangerous.

At this point this framework is still underdeveloped. For instance, I have not given you any reason why this way of identifying propositional attitudes with predictions is supposed to work for all attitudes. I also did not provide any proposals for the margins of propositional attitudes. If different predictive profiles can be identified with a folk psychological propositional attitude, then we ought to say how much the profile can change without changing the folk attitude. My description stayed on a general level, still in need of being fully spelled out. However, for the purpose of this paper it is sufficient to provide a rough idea how we need not rely on more than two different modes of predictions, both solely distinguished by their direction of fit, and precision-weighting. The important point is to provide an attempt of an account of self-knowledge in a predictive processing framework, and we now have enough resources to transform the double bookkeeping picture for perception and experience into a more general version for perception and mental states.

To label one part ‘perception’ and the other introspection is slightly misleading, because the same engine is at work in both. Therefore, I am going to speak of first-order processing and second-order processing for the two ‘books’ in Hohwy’s analogy. First-order processing relates to states about the world, second-order to states about one’s mind. The basic idea is the same as in the model for expectations of experiences. The important difference is that now these are replaced with expectations of predictions on the next lower level. For instance, Layer 2 on the first-order processing strand predicts both a sensory input from Layer 1 and a prediction on MLayer 1 (together with the next higher MLayer). If everything

goes right the second-order prediction on MLayer 1 matches the sensory input prediction on Layer 2. In other words: Based on Layer 3 the brain produces a $belief_{pred}$ of certain sensory input on Layer 2, and (together with the next higher layer) a $belief_{pred}$ that it $believes_{pred}$ a certain sensory input on Layer 2. Layer 2 sends prediction errors upward to both, either, or neither Layer 3 and MLayer 2.

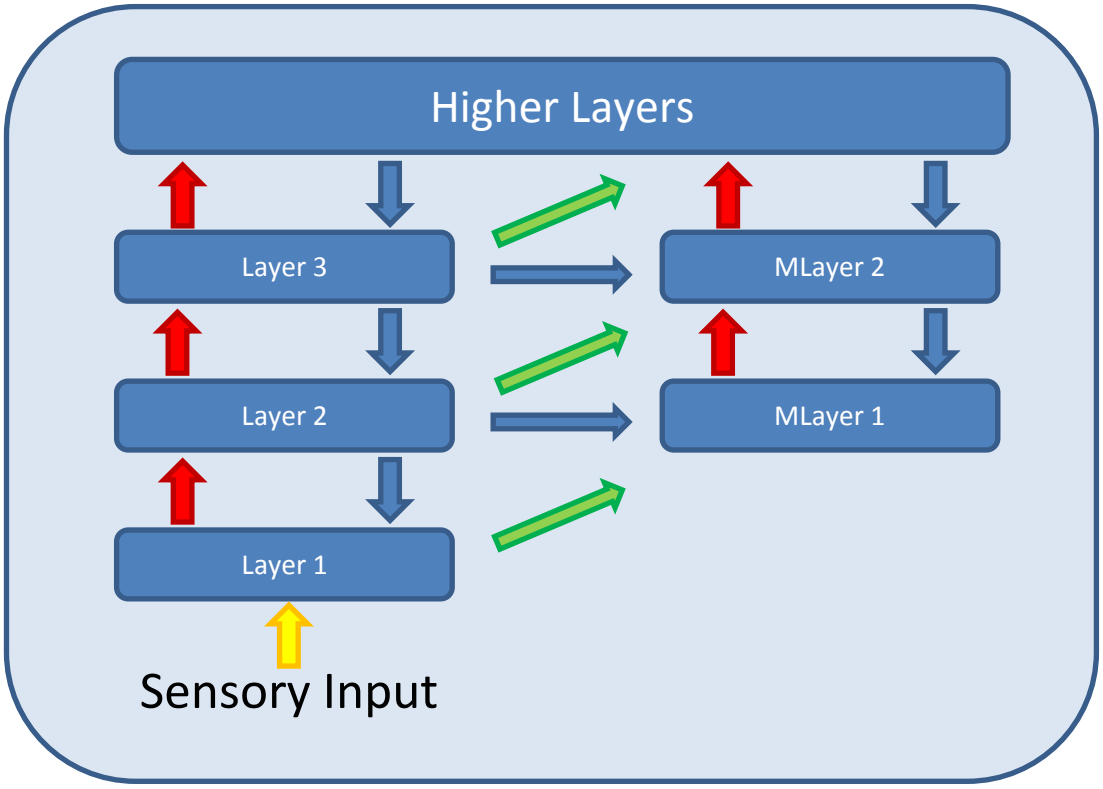

Figure 9

A difficulty arises when we simply substitute expected experiences with expected predictions. The experiential strand had the neat property that it was possible to compare predicted experience with actual experience and then feed the respective prediction error upwards. However, the model does not have the capability to compare predicted prediction to actual prediction. The prediction error that is sent upwards by, say, Layer 1 is only the difference between the prediction and the sensory input received at Layer 1. Moreover, if we want to avoid any form of inner-sense model of introspection for attitudes, the model better not involve a mechanism that directly compares second-order predictions with first-order predictions, because this would be a mechanism that checks whether the second-order $belief_{pred}$ fits with the first-order state and hence requires a way to access the complete first-order state.

There might be a conceptual way around this problem. So far I presented the second-order predictions as predictions of predictions. However, we can think about them slightly differently in terms of predictions of an activity. That is, based on, say, Layer 2 the brain predicts that it is going to predict a certain sensory input at Layer 1. This prediction of an activity is captured in MLayer 1. This is not a switch in the notion of prediction in play. There is no relevant difference between 'prediction of prediction' and 'prediction of predicting.' Both are spelled out in exactly the same way in the predictive processing framework, i.e. as predicting certain neuronal activity. However, this different point of view gives us the conceptual option that predicting an activity might also be a case of predicting a certain movement of one's body – predicting one's action. It is a prediction of the way in which one is going to minimize the error in Layer 1. Predictions of one's activity in this sense can be compared with one's body's actual movement. Hence a prediction error can go upstream.

The big question here is then how this prediction error can relate the second-order predictions to the first-order predictions. My proposal here is the following: active predictive processing allows for the brain to secure the correct fit for its predictions about the world in two ways: Either by changing the model and thereby the prediction, or by changing the world and thereby acting. Any Layer in the first-order strand is capable of eliciting action. These acts can be compared with the predicted action based on the second-order strand to generate prediction error for the second-order strand.

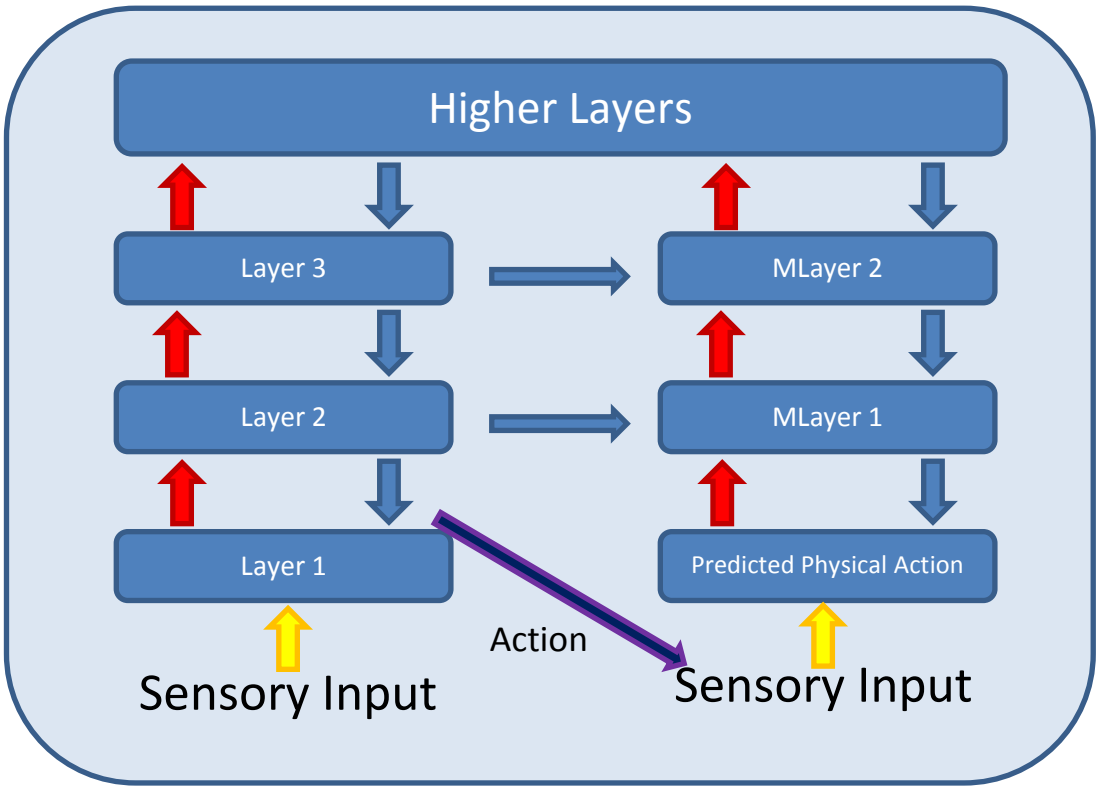

Figure 10

Comparing the predicted action and the actually performed action does not require any direct access to first-order mental states, so we can avoid falling back into an inner-sense view of introspection. Importantly, predicting an action is also nothing else but predicting a certain sensory input. Moreover, because the second-order processing ends up with a prediction of a certain sensory input and this can be compared with the sensory input that actually arrives, we can describe how second-order processing can be trained. It can be trained because whenever there is no, or little prediction error on the first-order processing side of things, while there is still a prediction error between predicted action (in terms of predicted sensory input) and actual action (in terms of sensory input) this can be fed upwards to change the second-order processing model. In case there is a prediction error in the first-order side then the prediction error on the second-order strand is irrelevant, because the problem might be inherited from the first-order processing.[120]

[120] However, it might be possible to train both sides at the same time. I want to remain neutral on this issue.

We can describe the brain according to this model as permanently keeping track of itself. It tries to predict the world, and itself. At both levels the brain improves its models by noticing whether the predictions have been erroneous.

6.5 Empirical Requirements, Predictions, and Support

Double bookkeeping does not imply double the channels of sensory inputs. All that is required is that the sensory inputs are processed in a way that generates two prediction errors that are fed forward. One related to the low level predictions of the incoming sensory signals in general, and one related to the sensory signals connected to one's own body in the world. To do this, sensory signals need to be understood not as merely input from outside (exteroception), but also include proprioception and interoception. Importantly, it would be a mistake to think of the prediction of actions (in terms of sensory input) only on a proprioceptive and interoceptive basis. Multiple experiments indicate that exteroceptive input is an important factor in locating one's own body and actions (Maravita, et al., 2003; Blanchard, et al., 2011; Guerraz, et al., 2012). Moreover, in some cases visual input seems to be the dominate factor (Lishman & Lee, 1973; Botvinick & Cohen, 1998; Drummer, et al., 2009). Therefore exteroception has to be included in the process generating prediction errors related to one's actions.

My proposed picture of double bookkeeping uses two prediction errors based on sensory input for the two different strands of predictions: first-order and second-order predictions. Either strand can be trained in virtue of prediction errors and change of the internal models underlying the predictions. If this is correct, we can expect to find indications of these trainings. Training the first-order processing is uncontroversial, because it only requires that the internal model of the world (or part of it) changes when the predicted sensory input does not match the actual one. Training in this sense is just learning about the world. I focus on training for the second-order predictions, which ultimately end in predictions of one's action (in terms of sensory input). Training here can only come from a mismatch between the predicted action and the actual action. That is a mismatch between the predicted sensory input of one's own bodily movement and the actual one. In other words, training of the model happens at moments of surprise. We can capture surprise in a quantifiable way by ignoring the phenomenal elements of it. Surprise in this sense becomes "the negative log-probability of an outcome" (Friston, 2010, p. 128). Following this definition lower subjective probability of an outcome comes with higher surprise if it

occurs. Tribus (1961) uses ‘surprisal’ instead of ‘surprise’ to mark this difference between the phenomenal surprise, and the quantifiable negative probability of an outcome.

Given that successful prediction requires minimization of surprise, we can expect that a brain that is still at an early stage of training its models (and hence its predictions) will be met with surprising events frequently, whereas a more refined, well-trained model meets less surprises. For the prediction of one’s own behavior based on second-order predictions this implies that we should expect cases in which one is surprised by one’s own behavior. Moreover, we should expect that how exactly these cases look like will differ from infants and young children to adults, because infants and young children predict in a coarser grained manner than adults (Baillargeon, 1994a; 1994b).

It is difficult to know when infants and young children are actually surprised by their own behavior. Methodologically it seems to be easier to locate this phenomenon in older children and adults who can report on the cause of their surprise. In the following I discuss a case of being surprised by one’s own behavior. This is the everyday, benign surprise of the ‘Broken Escalator Phenomenon.’ This also has the advantage of being an easily accessible and relatable case. Even though I will look at an experimental setting, the phenomenon is one with which most of us are familiar with.

Reynolds and Bronstein (2003) studied motor aftereffects by using a combination of fixed platform and mobile sled. Subjects were trained by moving from the fixed platform onto the mobile sled for 20 trials. After the moving trials subjects were given clear, verbal warnings that the sled would keep stationary for the next trials. Subjects then were asked whether they heard and understood the warning before walking another 10 trials. Body positioning and walking velocity were recorded for all trials. This setup is not too different from other, previous motor aftereffect experiments tracing back at least to Held (1965). The interesting part of Reynolds’ and Bronstein’s experiment is that they included reports of the subjects experiencing the effect. The bodily result of their experiment is unsurprising to anyone familiar with broken escalators. After about five trials with the moving platform the subjects adapted to the movement. Their velocity increased and soon after their posture changed as well, leaning slightly forward. The trials on the stationary sled afterwards showed that they still compensated for the moving sled, even though they reported full awareness of the sled being stationary. Moreover, they reported being surprised by their own behavior. Reynolds and Bronstein write:

> Most subjects spontaneously expressed great surprise and amusement when, on walking on the stationary sled, the aftereffect occurred. When subsequently questioned, however, all confirmed that they had understood and believed the experimenters' warning that the sled would not move. Those subjects who could relate to the broken escalator phenomenon found the experimental aftereffect to be similar to the real-life experience (Reynolds & Bronstein, 2003, p. 305).

Subjects are surprised by their own behavior. They claim to know that the sled is not going to move but nevertheless they behave as if they were stepping on a moving sled. Reynolds and Bronstein interpret this as the motor system acting inappropriately even though perception and cognition are veridical (Reynolds & Bronstein, 2003, p. 306). They generalize and propose an explanation in terms of "dissociation between declarative and procedural systems in the central nervous system" (Reynolds & Bronstein, 2003, p. 308).

How does this relate to my proposed picture of double bookkeeping? First a note of caution: There is a danger of identifying the quantifiable, technical use of surprise (the surprisal) with the phenomenal experience of being surprised. There might be a disconnection between surprise (surprisal), in the sense of a mismatch between predicted, and actual sensory input, and experienced surprise. Clark (2013) gives the example of an elephant on a magician's stage. One might be surprised by the elephant, even though the brain gives this state of the world a high probability (low surprisal). However, both senses of surprise are easily reconciled. Even though the elephant might be predicted as quite probable now, it wasn't a moment ago before it showed up. Prediction error from sensory inputs caused a revision of the internal models that then in turn have the elephant as a probable hypothesis (Clark, 2013, p. 16). It is plausible that the phenomenal surprise is therefore caused by an initial surprise (surprisal) on the prediction level. It then may last, even though the mismatch is resolved. This is enough to infer some suprisal from reported surprise, and allows us to use the relatable case of the broken escalator to illustrate how one assesses one's own mental states.

The subjects being surprised in both the phenomenal and technical senses marks a mismatch of the predicted behavior (action) and the actual behavior. However, it does not seem to be all that clear whether perception and cognition are veridical. Rather, what we can know from the reports of the subject is that reports of perceptual beliefs and cognition fit with how perception and cognition ought to be in the subject's situation. Subjects report

that they believe the sled is stationary. By these means we cannot know whether subjects really believe$_{prop}$ it is stationary. Their behavior indicates on the contrary that they do not believe$_{prop}$ it is stationary. If we rate the behavior as a better indicator than the report, then what we find here is a mismatch between belief$_{prop}$ and second-order belief$_{prop}$. Given that beliefs$_{prop}$ entail beliefs$_{pred}$, we have a mismatch between prediction and second-order prediction. This mismatch is then iterated in the next step insofar as the actual action and the predicted action do not match.

In this picture the mismatch starts when at the first-order strand a moving sled is predicted, while at the second strand a prediction of a stationary sled is predicted. The former is based on the previous experiences in the setting, the latter also on the testimony by the experimenter.[121] When the subject declares that she knows the sled is stationary, she does so in virtue of the prediction that she predicts a stationary sled.[122] That is, she declares in virtue of her second-order belief. This mismatch is then reiterated on the lower levels. However, the mismatch need not be only in terms of content. Both strands fit insofar as one predicts input based on steady movement, and the other expects a prediction based on steady movement. The difference here has to be in the direction of fit of the prediction. While the steady movement prediction on the left side is based on a moving sled, the second-order prediction on the right is based on the second-order prediction of predicting a stationary sled. The impact is that on the left side the aim to reduce prediction error results in an action to counteract the moving sled. Steady movement on a moving sled requires changes in velocity and posture. The final left slide prediction is therefore a world-to-mind directed one, such as desire, or intention. On the right side the expected prediction of steady movement is based on the expected prediction of a stationary sled, hence no need for an action is present. Here we have a mind-to-world direction of fit – a belief.

[121] A different interpretation is that both bases include the testimony, but the weighting of the testimony differs.

[122] There might not be a single prediction that captures the stationary sled. Instead this state of affairs will be represented over multiple predictions on different layers. For simplicity I treat this as the content of a single prediction here. The same goes for all other predictions in this description.

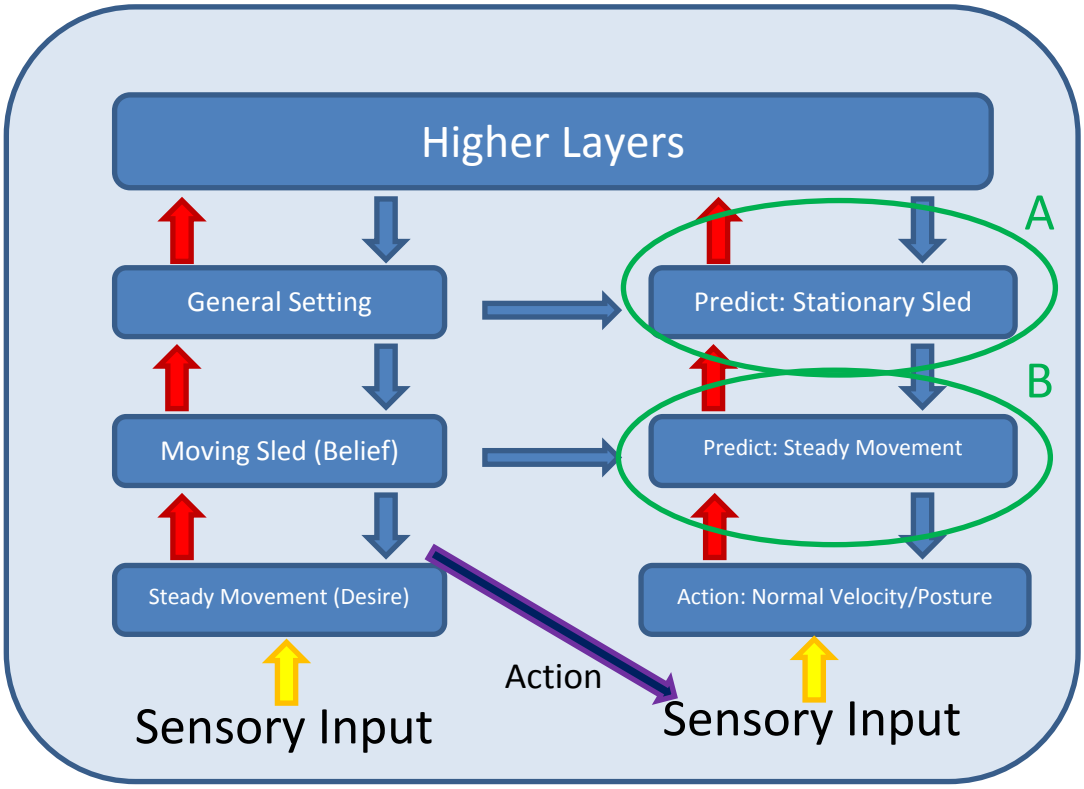

Figure 11

When Reynolds and Bronstein talk about perception and cognition being veridical they point to the fact that the predicted stationary sled belief (A) and the predicted steady movement belief (B) are exactly what subjects should have in the situation. However, their actual perception might not be veridical. Rather, they perceive the situation as if there was a moving sled, which causes subjects to counteract the moving sled. They are then surprised by their own action, because the action is based on a moving sled, whereas the predicted action is based on a stationary sled. This still fits well with the general explanation as "dissociation between declarative and procedural systems in the central nervous system," (Reynolds & Bronstein, 2003, p. 308) if we identify the declarative system with the right (the second-order) strand, the procedural system with the left (the first-order) strand.

Surprise in this case is not limited to the subject's own behavior. Surprise also occurs as a result of the overcompensation of the assumed moving sled. The brain predicts a steady movement and wants to match this prediction with the sensory input by performing a certain action (moving faster, changing posture). It turns out this action increases

prediction error, so the internal model has to change to adjust the prediction. After enough iteration the internal model will be on a stationary sled model, in which case the actual action matches the predicted action. There is no need to change the second-order models if a prediction error at the first-order side was reported. However, we can expect different cases in which there is no or no significant prediction error on the first-order side, but a significant prediction error on the second-order side, based on a mismatch of action and predicted action. In this case the models of the second-order side have to be changed to minimize prediction error.

6.6 Explaining Self-Knowledge

I can now formulate my proposed predictive processing story of self-knowledge for propositional attitudes: self-knowledge is based on second-order predictions (mind-to-world directed). You believe$_{prop}$ that you believe$_{prop}$ that p, if you expect yourself to have a certain predictive profile. And you expect yourself to have a certain predictive profile in virtue of multiple second-order predictions, each single one a prediction of a prediction or action. Second-order predictions are formed at the same time as first-order predictions, which is precisely the idea that I captured in the single process model. There are not two distinct ways of forming first-order belief$_{prop}$ and corresponding second-order belief$_{prop}$. They are both produced by the same machinery at the same time. Furthermore, you can only know that you believe that p, if (i) your process of generating second-order predictions is generally reliable[123], and (ii) your current, relevant second-order predictions are largely correct[124]. Largely, because it does not seem necessary that every single second-order prediction is correct. It is conceivable that a single second-order prediction does not match any first-order prediction, but the pattern of second-order prediction nevertheless realizes a second-order belief$_{prop}$ that matches a first-order belief$_{prop}$. There seems to be some margin of error. This fits with the undemanding fictionalist stance that merely requires that our notion of belief$_{prop}$ allows us to make predictions and explanations of behavior. As long as this function is fulfilled beliefs$_{prop}$ can be imprecise. Propositional attitudes therefore need not allow us to infer the exact predictions in place. Moreover, different sets of predictions might instantiate behavioral dispositions that are similar enough for our purposes in folk psychology, which gives us further reasons not to attempt to infer specific predictions from the ascription of a propositional attitude.

[123] i.e. 'justified' in the reliabilist sense
[124] i.e. your second-order belief$_{prop}$ is true

The reliability of second-order predictions is established by a feedback loop based on expected actions and actual actions. This way the account gives us an explanation of why we can accurately ascribe our own attitudes. However, the account also has room for fallibility. It is possible that one's second-order predictions are wrong and that one's prediction of an action is mistaken. Moreover, the framework provides room for such mistakes in self-knowledge to be consistent over time. Suppose that one predicts a specific action based on a series of second-order predictions. Further suppose that this action does not actually occur. The brain then generates an error signal based on the mismatch of predicted action and actual action. In these cases the brain always has to decide which one it should trust, the priors which generated the prediction, or the incoming prediction error. The mistake may be located in either of those. In a usual case the incoming sense data that is responsible for the prediction error will be trusted and the priors adjusted accordingly. The model recalibrates itself based on this feedback loop. However, in some cases the priors might have such a high assigned probability that it seems more likely that the prediction error itself was a mistake. In that case, the prediction error will be disregarded and the priors stay as they are. Remember Peacocke's (1998) case of the biased administrator:

> Someone may judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (Peacocke, 1998, p. 90).

The administrator takes herself to be fair and unbiased. We can suppose that her brain has second-order predictions that culminate in a prediction to act in a manner that treats all applicants the same. If she now receives a mismatching sensory perception that can be interpreted as treating applicants from other countries worse, her brain has to make a choice (so to speak): was she wrong about her predictions of herself, or is the reported error a false positive? If the latter choice is deemed more probable, then there will be no change to her priors and predictions. She will still take herself to be fair and unbiased, even though her actions tell a different tale. The persistence of the undetected bias is explained insofar as actions that do not fit with her view of herself are not entering the feedback loop that impacts her predictions. Instead, these error signals themselves will be reinterpreted to fit with her priors.

For a full account we need to combine this with an explanation for knowledge of experiential states, and knowledge of non-propositional attitudes. If my proposed model can work for propositional attitudes, then there seems to be no good reason why it should not for non-propositional attitudes. On the contrary, it should be easier to achieve this, insofar as predictions are usually taken to be non-propositional. Believing that I fear a spider would be a pattern of second-order prediction, just as believing that I believe that 'this grass is green' is. For experiential states Hohwy's (2013) explanation is a starting step, but needs to be spelled out in more detail – a project I leave open here.

However, if this proposed cognitive model turns out to be correct, the main lesson we should take away from it is that forming second-order $\text{beliefs}_{\text{prop}}$ ought to be understood as part of the same process as first-order $\text{belief}_{\text{prop}}$-formation. Second-order predictions are not formed independently of first-order predictions. Rather, they are connected downward and sideward. All part of a single machinery trying to predict the world and itself. We should not think of introspection as a distinct belief-forming process that detects our mental states, but as a built-in part of our predictive processes. Self-knowledge can thereby be at least partially based on the very same thing that is the basis for first-order states. The proposed predictive processing model therefor qualifies as a version of the single process model and as a transparency account of self-knowledge. It proposes that a mental and knowledge of that mental state can be formed based on the same outward phenomenon (Evans, 1982). Outward phenomena in the predictive processing case are predictions (in either direction of fit) about the world and the error signals, both ultimately based on sensory input.

Furthermore, the cognitive proposal of the single process model can also account for the privileged nature of self-knowledge compared to knowledge of other people's mental states. Predictions of predictions and predictions of actions are available differently for oneself and for other persons. For other persons I only have access to observations of their behavior which can be a basis for predicting their mental states (mind-reading). In contrast, my brain has access to fine-grained first-order predictions as a basis for second-order predictions. This availability of sub-personal predictions accounts for the high accuracy of self-knowledge, and the peculiar nature of self-knowledge. Because these predictions are made on a sub-personal level self-knowledge appears immediate or groundless. I do not have conscious access to the basis of my second-order $\text{belief}_{\text{prop}}$, because I have no

conscious access to the basis of any involved $belief_{pred}$. Notice that the basis for the second-order $belief_{prop}$ is not the first-order $belief_{prop}$, but rather the individual first-order $beliefs_{pred}$ and relevant higher level second-order predictions. This is the result of treating the $belief_{prop}$ as a pattern of $beliefs_{pred}$. The $belief_{prop}$ is not taken to be a single entity. Instead it is a group of $beliefs_{pred}$ taken together that we identify it with our folk notion of belief in a fictionalist manner.

6.7 Conclusion

I provided a predictive processing story that fits the single process model for self-knowledge. I proposed that we should understand self-knowledge of mental states based on a double bookkeeping picture of predictive processing. Two connected predictive strands run in parallel, one involving predictions about the world, and the other one involving predictions about predictions or one's actions. These predictions can come in two directions of fit: mind-to-world and world-to-mind. I suggested that we should identify our folk notions of propositional attitudes with patterns or profiles of predictions in merely a fictionalist manner to keep a connection between our ordinary talk and our study of cognition. If we do so, we can accept second-order propositional beliefs as capturing a set of second-order predictions. Hence, you $believe_{prop}$ that you $believe_{prop}$ that p, if you expect yourself to have a certain predictive profile. And you expect yourself to have a certain predictive profile in virtue of multiple second-order predictions, each single one a prediction of a prediction or action. Furthermore, you can only know that you believe that p, if your second-order predictions are largely correct and the process generating them is reliable. This picture explains the nature of self-knowledge as peculiar and privileged, because this basis for your second-order predictions is different from the observational basis of beliefs formed by mind-reading. The model is still incomplete. We require a better understanding whether the proposed relation between propositional attitudes and non-propositional predictions is plausible and how second-order predictions fit into child development. I merely provided a first step towards a cognitive story for the single process model based on the predictive processing framework.

7 Extending Introspection

Clark and Chalmers (1998) propose that the mind extends further than skin and skull. If they are right, then we should expect this to have some effect on our way of knowing our own mental states. If the content of my notebook can be part of my belief system, then looking at the notebook seems to be a way to get to know my own beliefs. Moreover, if the single process model aims at being a unified account of self-knowledge, I ought to say something about knowledge of extended mental states. However, it is at least not obvious whether self-ascribing a belief by looking at my notebook is a case of introspection the same way that knowing my non-extended beliefs is. Traditionally this sort of introspection is thought to be privileged and special in ways that the extended introspection case seems not to be. There is nothing privileged about looking at my notebook. Anyone could do it. The aim of the chapter is to find out how to understand extended introspection and whether there is something privileged and special about knowing one's own extended beliefs. First, I present the case of extended introspection. I then discuss whether it should be understood as genuine introspection or as mind-reading. Both seem to be bad fits, which finally prompts an original account for extended introspection based on epistemic rules.

7.1 Introduction

I started out with the aim of providing a unified transparency account of self-knowledge. So far I showed how the single process model explains self-beliefs (self-knowledge in a good case) for attitudes and experiential states and provided an attempt at a cognitive story that fits the account and empirical evidence. However, when discussing mental states I only discussed cases in which these states are bound to the limits of the brain and skull. Clark and Chalmers (1998) propose that there are other mental states that are not fully located inside the brain. Some mental states are extended to external devises. A truly unified account should also tell us how we can know our own extended mental states. Take the following case of extended introspection:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. In short, Otto has beliefs extending to the notebook. One day I meet Otto in the streets of New York. Out of curiosity I ask him "Otto, I know where the museum is, but do you believe the museum is on 53rd street?" Otto looks at his notebook, finds the right entry, and answers "I believe the museum is on 53rd Street."[125]

[125] This is different from the extended self-knowledge case by Carter and Pritchard (Forthcoming), which has the self-ascription written down in the notebook: "[…] For example, when he learns new information about his own mental states (i.e., beliefs,

At face value in this story Otto self-ascribes a belief state by looking at the notebook. But this leads straight into conflicting intuitions:

- On one hand Otto appears to simply detect his belief. The story is set up in a way that he has beliefs extending to the notebook, so the way to detect these beliefs is to look at the notebook. And intuitively, what else is introspection other than directly detecting your own beliefs?
- On the other hand Otto appears to base his self-ascription on evidence that I, standing next to Otto, can use just as well to ascribe the belief to Otto. And this is certainly not what we intuitively think introspection is like.

With these conflicting intuitions at hand, we need to search for a proper way to understand extended introspection. Otto definitely self-ascribes a mental state by looking at the notebook. But what kind of self-ascription are we dealing with here? This is the core question I am going to address in this chapter. I start by stepping back and offering a quick overview of the notion of extended belief as the basis for extended introspection. I then consider the standard options for self-ascribing mental states: introspection, and self-directed mind-reading. I discuss these in turn and show that neither are a great fit. This leads into a dialectic dilemma with no clear way out. I then propose a unique account of extended introspection based on epistemic rules, inspired by Alex Byrne's (2005) account of introspection. On my proposal extended introspection turns out to be reliable, because self-verifying under extended belief conditions.

7.2 Extended Belief

My story about Otto introspecting extended beliefs builds on the Otto case presented by Clark and Chalmers (1998). The central idea behind ascribing Otto plus notebook an

---

feelings, desires, etc.) – information about his mental states which would be lost in biological storage – he writes it down in the notebook. Likewise, when he needs some old information about his mental life, he looks it up. For Otto*, his notebook plays the role usually played by a biological memory in preserving a mental narrative." I will use 'extended introspection' instead of 'extended self-knowledge' to avoid confusion.

extended belief lies in the fact that the notebook plays the role usually played by biological memory. This idea features in the argument as the parity principle:

> *Parity Principle*: If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process (Clark & Chalmers, 1998, p. 8).

The notebook does the job of the biological memory, because Otto consistently uses it just like we usually use a biological memory. Whenever he gets new information he writes it down. Whenever he wants to remember something he looks it up in the notebook. And the notebook is with him all the time. So why not straight up accept the notebook as Otto's memory?

The difference in location is not a well-motivated rationale to dismiss the similarities, if we follow Clark and Chalmers. However, this needs a further, more detailed look. Even if we follow the parity principle, we need to say something more on the notebook's role in Otto's cognitive processes. Not every notebook makes a good case for something that "we would have no hesitation in recognizing as part of the cognitive process," that is in this case, as part of memory. I personally use notebooks every once in a while. But if you were to observe my sporadic looks at the notes, you would surely hesitate to call my notebook part of my memory. Moreover, sometimes I write notes so illegibly that there is no way to decipher the content later on. Other times I find myself disagreeing with my notes – "I can't have meant that," I say to myself. And finally, often my notebook is just not with me, whereas my biological memory is consistently with me. Or at least I cannot remember the cases in which it was not. In short: My notebook's role is very differently from biological memory. We have to be careful not to trivialize extended beliefs in such a way that too much satisfies the criteria.

Clark and Chalmers are fully aware of this concern. Therefore they propose four conditions that external aids have to satisfy to play a role in an extended belief (Clark & Chalmers, 1998; Clark, 2010).

External aids must be

(i) consistently available,

(ii) readily accessible,

(iii) automatically endorsed, and

(iv) present because they were consciously endorsed in the past.

They are slightly skeptical about the previous endorsement condition, claiming that this requirement is debatable. The appeal of (iv) is that by getting rid of it one might lose grip of the difference between remembering and relearning. However, insofar as (iv) requires conscious endorsement, which is an internal condition, (iv) seems to rob the extended belief thesis of its core motivation. The cognitive work would not be extended enough, so to speak (cf. Bernecker (2014, p. 5)). For my purposes condition (iv) is not important, so I leave the debate open. However, conditions (ii) and (iii) will become important later on. Fortunately, they are generally accepted as requirements for extended beliefs.

Clark (2008) and Menary (2007) introduce additional conditions of two-way interactions. The idea is that the interactions between the external aid and the person have to be connected in such a way that they continuously affect the other. Clark calls this *continuous reciprocal causation*, Menary *cognitive integration*. A simple book, for instance, would not fit this criterion because the reader does not affect the book sufficiently. For my purposes the conditions (i) to (iii) above are enough, so I can stay neutral on any further requirements. I will also remain neutral on whether the analogy to memory works as well as Clark and Chalmers want it to. The use of notebooks can fail more frequently and in different ways than our biological memory, which might undercut the parallel drawn by the argument for the extended mind (cf. Rupert (2004), Sterelny (2004)). However, for the present purpose of discussing extended introspection, I will assume that the case can be made in favor of Clark's and Chalmers's picture. Hence, my approach can be seen as a discussion under a conditional: What should we think of introspection of extended beliefs, *if there are extended beliefs and they can be roughly characterized by conditions (i) to (iii).*

There is a final worry to address concerning the focus on the simple Otto case. The case is highly idealized. Why should we care about this case at all? Perhaps looking at other, more realistic scenarios requires different conditions or dimensions of integration (as discussed in Sutton (2006), Sutton et al. (2010), Sterelny (2010), Menary (2010), and Heersmink (2015))? My first response here is to point out that the case does not strike me as unrealistic or abnormal, even though it involves an uncommon situation. It seems similar enough to the use of notebooks, smartphones, or smartwatches as storage for information in our

everyday practice. Moreover, because the simple case only uses technology that is well understood it has a clear advantage for pumping intuitions about possible extended beliefs: We are well acquainted with the imagined situation and relevant counterfactuals. In contrast, using more speculative scenarios, such as cognitive enhancements directly connected to the brain, demands us to conceive of something rather alien to us. This often has the result that we cannot make judgments about these scenarios with high confidence. We simply do not know enough about how these cognitive enhancements would work in practice. There is no such problem with the simple notebook case. Finally, even though the simple case might not represent all cases of extended belief, it gives us a good starting point for theorizing. We need to understand the simple cases properly before tackling the more complex ones and, at least insofar as we want to hold on to the parity principle, (i)-(iii) seem to be a good way to capture extended belief in these simple cases.

7.3 Extended Introspection as Introspection

To find out whether the extended introspection case is a genuine case of introspection, we need to define our criterion for such genuine introspection. A good starting point for doing so is to look at our intuitive judgments about introspection, and most importantly what they are contrasted to. As discussed chapter 1 the two important points of contrasts are first, knowledge of the non-mental world; and second, knowledge of other people's mental states. Introspection appears to be different to both. Self-knowledge seems to some extent *privileged* and *peculiar* (Byrne, 2005). It is privileged insofar as one is more likely to know one's own mental states than other's mental states or the external world. Moreover, they are peculiar, insofar as they are formed by a special method or way of knowing. Call this the *cognitive access view* of self-knowledge. Cognitive access accounts come in different shapes. For instance, they can accept a peculiar detectivist method of introspection (e.g. Armstrong (1968), Nichols and Stich (2003), Goldman (2006), Macdonald (2014))[126] or an empiricist transparency story as in Byrne (2005), and Fernández (2013).

The cognitive access view is not the only description of self-knowledge available. In chapter 1 I discussed s*elf/other parity* accounts of self-knowledge, which argue that the specialness of self-knowledge is overstated. Self-beliefs are largely[127] formed by the same processes we

[126] These lists are not exhaustive.

[127] 'Largely' because they usually limit the parity to propositional attitudes. Moreover, Carruthers accepts that the parity does not fully hold with regard to knowledge of one's phenomenal states because generating knowledge of these states does not require interpretation.

use to attribute mental states to others (cf. Carruthers (2011), Cassam (2014)). Moreover, I pointed to a*gentialist*[128] positions (e.g. Burge (1996), Moran (2001), Bilgrami (2006)) that understand privilege and peculiarity not in terms of better access, but with a particular first-personal connection between (rational) agents and their mental states. For Moran (2001) this particular connection is also the basis for the transparency of beliefs, that is, that one can know whether one believes that p by attending to the question of whether p is true. Agentialist accounts accept that self-belief is special because it is about *my* mental states, and I am responsible for *my* beliefs. Self-knowledge in this conception is important as a precondition for critically reflecting on one's own mental states.

I chose the cognitive access view earlier (chapters 1, 2 and 4) and will restrict myself to the cognitive access view for extended mental states as well. Hence, I understand privileged access for introspection as a person being more reliable in self-ascribing mental states than in ascribing mental states to other people. Moreover, I take peculiar access to denote some sort of peculiar way of knowing one's own mental states, compared to knowing other people's mental states, or knowing one's own mental states by inference and interpretation (I call both of these options 'mind-reading' and discuss them in more detail later).

With the features of privileged access and peculiar access as the defining features for introspection we can start looking at whether extended introspection presents us with these features. When Otto self-ascribes his belief by looking at the notebook, does he have privileged and peculiar access? If the answer is negative, then the single process model will not be able to account for knowledge of all possible mental states. However, neither will any other account of introspection, so the single process model is in no worse position than its rivals.

The first step to investigate his privileged position is to consider whether Otto is reliable. Clearly he is in the way the story is set up. Otto has, by stipulation, the extended belief that the museum is at the 53rd street. When I ask him, he looks at the notebook and avows that he believes the museum is at the 53rd street. And in doing so, he makes a correct statement. It is true that he has this belief. Even more so, it seems very difficult for Otto to be wrong about himself in this case. Given that the notebook is consistently available,

[128] Sometimes called *rationalist* positions (Gertler, 2011a).

readily accessible, and automatically endorsed, Otto is highly reliable in looking at the notebook and self-ascribing the belief. In nearly all nearby possible worlds in which Otto looks at the notebook to self-ascribe the belief, he will read the notebook correctly, understand what is written in the notebook and successfully attribute the belief to himself. If this was not the case, then Otto would fail to have the appropriate connection to the notebook and not have an extended belief. But I stipulated the extended belief conditions to be satisfied. Hence Otto has to be reliable.

However, is Otto *more* reliable than other people looking at his notebook? Privileged access requires more than just reliability. It requires being more reliable than other people in attributing mental states to Otto. Consider the following take on Otto:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. I look at Otto's notebook and read that the museum is at the 53rd street. Therefore I judge that Otto believes the museum is at 53rd street.

Is my belief that Otto believes the museum is at 53rd street reliably formed? This depends on how we individuate the process. If the process in question is "looking at notebooks" in general, then it probably is not. However, given that I stipulated Otto's extended belief, it seems the answer is a clear yes if we individuate the process based on looking at *Otto's* notebook. That is, if I ascribe Otto beliefs based on looking at *his* notebook, I will be highly reliable. As long as I can read Otto's handwriting and understand his language, I will end up with true beliefs about Otto's extended beliefs.[129] Therefore I seem to be equally reliable as Otto in ascribing extended beliefs to Otto. Otto is not privileged.

Perhaps I am a little too quick here. One might bring up at least two objections. First, one may claim that there is a sense in which Otto is privileged. Otto has his notebook with him all the time, whereas I don't have the same kind of permanent access. So he has better access after all! To this I respond that this is not the right sort of privilege that I am talking about. It is important to distinguish two kinds of privileged access here. First, one can have 'accidentally' privileged access because one is more often in possession of the relevant evidence, but that evidence (or sufficiently similar evidence) is in principle accessible for

[129] And there seems little reason to give Otto any privilege with reading and understanding the notebook, or at least nothing that could not be removed by modifying the case slightly.

other people. Second, one can have 'essentially' privileged access, when the evidence (or sufficiently similar evidence) is not in principle accessible for other people. Otto seems to have accidentally privileged access to his extended belief, but no essentially privileged access. And it is the latter that is of interest here. If I were to follow Otto all the time and look at his notebook constantly there would not be an extended belief of his that he knows but I do not.

Second, one can argue that one needs to have reasons to base judgements of Otto's beliefs on the notebook. One needs to know not only that this is Otto's notebook, but also that Otto is related to the notebook in a way that satisfies the conditions for extended belief. I can accept the former, but deny the latter. One needs some reason to ascribe a belief based on the notebook, but knowing that this is Otto's notebook is enough.[130] After all, Otto himself does not need to know that the extended belief conditions are met. He just needs to look at the notebook and self-ascribe the belief. The same applies to me ascribing mental states to Otto by looking at his notebook. I need to have some reason why this notebook relates to Otto's beliefs, but those reasons need not entail any of Otto's beliefs. These reasons are only necessary to prompt me to ascribe a belief on the basis of the notebook. They need not be reasons that guarantee truth of my ascription.

What about peculiar access? Is there anything special about Otto's self-ascription? At first glance what Otto does and what I do when we ascribe a belief based on the notebook seems not that different. We both look at the very same thing and use it as a basis for a belief ascription. It is the same notebook. Sure, he has access to the notebook constantly, but that does not change the fact that when we both look at it to ascribe a mental state, we do the same thing. Otto might use the notebook for much more, but for this single belief-forming process it is hard to find a difference. That the notebook constitutes his first-order belief does not influence the second-order belief formation. Moreover, there is no reason to assume more inferential work being done by me than Otto. Sure, I need to recognize that this is Otto's notebook. But so does Otto. I don't see any reason why I would need any additional inferential step that Otto does not need. I can treat whatever is written in the notebook as direct evidence for Otto's belief, just as Otto does. Gertler (2007) uses this

[130] This includes knowing something about the function of a notebook. Some knowledge of the function is necessary to distinguish one's notebook from a scrap of paper lying around. That it is my notebook and not a random scrap of paper provides a prima facie reason to attribute to myself the content. Thanks to Grace Helton for pointing out this issue.

observation to argue against the existence of extended beliefs. As I will show later this is not the route that I want to take. However, Gertler is right that this symmetry between Otto and me is bad news for peculiar access and overall bad news for extended introspection as a subspecies of genuine introspection.

7.4 Extended Introspection as Mind-Reading

Mental state ascriptions are commonly assumed to be either introspectively formed, or based on mind-reading.[131] I just argued that extended introspection does not fit the criteria of privileged and peculiar access, so it does not look like it goes into the introspection category, regardless which account of introspection we accept. One should then expect a better fit with mind-reading. And if so, calling it 'extended introspection' might actually be misleading. To see whether this is true I want to start with a rough and ready characterization of mind-reading. I understand mind-reading here as a capacity to attribute mental state to human beings based on behavioral observations and evidence of the situation/environment. To illustrate this take this simple mind-reading story by Jordi Fernández (2013, p. 4):

> Suppose that one of the things that you believe about me is that I want Barcelona FC to win the UEFA Champions League. Suppose, furthermore, that your belief is justified. What could justify your belief? Perhaps you heard me express that desire, or you observed me screaming at the TV while we watched one of the Champions League games, or you noticed my mood when I read in the news that the team was not doing so well in that competition (Fernández, 2013, p. 4).

Whatever can justify your belief about his mental state has to be something you observed. You cannot directly access his mental state, but rather you need to base it on the evidence you gather by perception. You can listen to his testimony and you can see him get emotional when watching the game. Perhaps even facial expressions showing his mood can be sufficient if you know enough about the situation he is in right now. Crucially, the mental state attribution is based on things other than the mental state attributed. Plausibly they are not completely unrelated – his emotional reaction is connected to the desire that Barcelona FC wins – but they are different things. This is not to say that mental states cannot play any role in mind-reading processes. Rather, the mental state that should be

[131] Some argue for a distinct method of knowing other people's mental states that is not inferential mind-reading (cf. Spaulding (2015)). I will not consider these options.

attributed at the end of a process cannot itself be the input of the very same process. If I ascribe mental state M_1 by process P_1 at t_1 I can later on use M_1 in process P_2 at t_2 to ascribe M_2. For instance, I ascribe the desire for Barcelona FC to win the Champions League to Jordi Fernández based on his behavior while watching a football game. Then later on I see him celebrating after the game finished. I can now use the previously ascribed desire to figure out that he is happy that Barcelona FC won.

Mind-reading approaches that start with behavioral observation plus evidence of the situation come in two varieties: as theory-theory accounts (e.g. Gopnik and Wellman (1994; 2012), Gopnik and Meltzoff (1997)) and as simulation accounts (e.g. Goldman (2006)). Both differ in what is done with the observational input, but for my purpose the focus is on the input itself. So I can work with a very simplified black-box model.

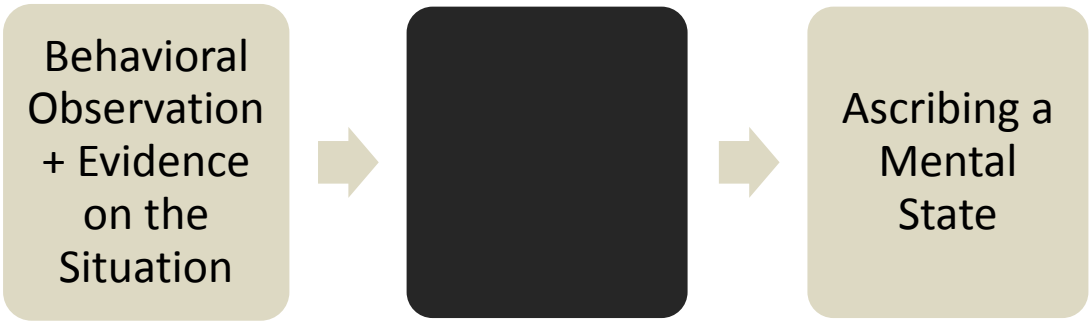

Figure 12

The model can be self-directed, so it is possible to self-ascribe a mental state based on behavioral observation and a grasp of the current situation. A standard example for this type of case is Wright's explanation of a scene in Jane Austen's Emma:

> Emma has just been told of the love of her protégée, Harriet, for her — Emma's — bachelor brother-in-law, a decade older than Emma, a frequent guest of her father's, and hitherto a stable, somewhat avuncular part of the background to her life. She has entertained no thought of him as a possible husband. But now she realizes that she strongly desires that he marry no one but her, and she arrives at this discovery by way of surprise at the strength and color of her reaction to Harriet's declaration, and by way of a few minutes' reflection on that reaction. She is, precisely, not moved to the realization immediately; it dawns on her as something she first suspects and *then* recognizes as true. It *explains* her reaction to Harriet (Wright, 1998, pp. 16-17).

With self-directed mind-reading in the mix, how can we tell a mind-reading story of Otto's self-ascription? Otto simply looks at the notebook and avows that he believes the museum

is at 53rd street. There is no strong, colorful reaction to the notebook that Otto then can interpret in such a way that makes it possible to attribute a belief. The only reaction is the assertion that he believes that the museum is at the 53rd street. But that reaction already presupposes what the mind-reading process wants to get at. It is useless as a basis for a mind-reading process. What other behavior can be considered as the basis for Otto? Perhaps he can base his mind-reading on previous instances of looking at the notebook. In the past he read the notebook and acted accordingly. But is this enough as a basis? I doubt it. While he might get to the general conclusion that usually he acts according to what is written in the notebook, there is no way this general claim can lead him to the specific attribution of a belief *that p*. Where should the propositional content come from, if his basis is a general claim?

The obvious amendment is to let the content written in the notebook play a role in the explanation. So the general observation of Otto's past behavior plus the fact that p is written in the notebook are the input for the mind-reading process. The output then is the self-ascription that he believes that p – that the museum is at 53rd street in this case.

But this is no good either. There are red flags for both parts of the input. First, it does not seem obvious whether Otto can use his past behavior at all as an input. Otto has Alzheimer's after all. How can we let any representation of his past behavior play a crucial role in the belief-production if it is unclear whether his memory supports any such representation? If Otto needs a notebook to remember where the museum is, he likely won't be able to remember how he behaved in the past. We can imagine him writing down all his behavior in the past as well, but that seems unnecessary in our story about Otto's self-ascription. The story is complete without him also checking the notebook for his previous behavior.

Second, if we accept that p written in the notebook plays a role as input, we might be already stepping away from mind-reading. Remember that the mental state attributed at the end of the mind-reading process cannot be already the input. If we let the notebook play such a pivotal role as input, we are in danger to abandoning this general rule, because the notebook stating that p is part of the extended belief that Otto wants to self-ascribe. This means that Otto would effectively use his extended belief to self-ascribe the very same belief. While this perfectly fits the story, it does not tell a mind-reading tale anymore. Instead we have to deal with Otto directly detecting his own extended belief. He bases his

second-order belief on his first-order belief. Suddenly the mind-reading approach points towards genuine introspection as the correct way to go.

The proponent of the mind-reading solution can attempt to save his account by implementing a different move. It is not the present behavior that Otto interprets, but rather his past behavior. That there is something written in the notebook is evidence of his own past action of writing down the location of the museum. When Otto recognizes what is written in the notebook, he recognizes that this is something that he wrote and hence, that he believes. In this case he bases his self-ascription on the evidence of his prior action. The first problem with this solution is that it heavily relies on Otto having written that p into the notebook earlier. However, as I noted before, the previous endorsement condition is not all that necessary according to Clark and Chalmers. And if there is no previous endorsement condition, then there can be cases in which the writing in the notebook is no evidence of Otto's past actions. Second, in some cases Otto will not be able to tell whether he wrote p into the notebook. Just think of notes on a smartphone. There is no handwriting that can be recognized as something Otto himself wrote. Nevertheless, he can have beliefs extended to the notes on the smartphone and be unable to rule out that someone else wrote them. It seems that he does not need to know who put the notes in there. If this is correct, then he cannot use these notes as results from his past actions. Hence, he cannot use them as a basis for self-directed mind-reading.

I showed that both the introspection and the mind-reading approach fail to capture extended introspection appropriately. What are we supposed to do now? I believe there are four options available, with little favor any of them.

1) We can change our account of introspection, such that privileged- and peculiar access are less important.
2) We can change our account of mind-reading, such that we can allow the ascribed belief to already be a part of the input in some sense.
3) We can get rid of the idea of extended belief altogether, and stick to our guns for introspection and mind-reading.

4) We can propose that extended introspection needs its own, distinct account of producing self-knowledge.

It is a difficult dialectic position to be in. Moreover, if one looks at the individual debates, then one can find independent motivation for every single of these options. One can accept that introspection does not come with this sort of privileged- and peculiar access in general with Carruthers (2011), Cassam (2014), or Schwitzgebel (2008) and go for (1). One can adapt ideas from direct social perception of mental states such as presented in Krueger (2012) or Spaulding (2015) and go for (2). One can get rid of extended minds with Gertler (2007) or Adams and Aizawa (2010) and go for option (3). I, however, want to opt for (4). The reasoning is largely defensive. If one chooses option (4), one need not abandon any mainstream position for introspection, mind-reading and extended beliefs. Given that these positions have something[132] going for them, (4) is the least invasive way to go. I can avoid the internal debates of introspection, mind-reading, and extended mind for the bargain of accepting a new source of self-knowledge: Extended Introspection.

7.5 Extended Introspection Sui Generis

Perhaps it should not surprise us that both introspection and mind-reading do not work for our story about Otto. Why expect a phenomenon to fit an explanation that was modeled after very different cases? So let's start looking at the case and build an account from the bottom up.

The main idea is to describe what is going on in the Otto case and then transform this description into a general principle or rule. This approach is not entirely original. Rather, it is a staple in the epistemologist's toolbox to build epistemic rules out of cases, whereas an epistemic rule is simply a rule of belief formation.

For instance, we can start with the following story taken from Alex Byrne (2005, p. 93):

[132] I am not going over the advantages of the mainstream positions in detail here. However, some of them can be seen in the fit with folk psychology and our linguistic practice. For instance, how they explain the peculiar nature of avowals, the phenomenology of introspection, and mental state disagreements, as discussed in earlier parts of the thesis.

> Mrs. Hudson might hear the doorbell ring, and conclude that there is someone at the door. By hearing that the doorbell is ringing, Mrs. Hudson knows that the doorbell is ringing; by reasoning, she knows that there is someone at the door.

This case is straightforward. Mrs. Hudson believes that someone is at the door, because the doorbell rings. We can transform this into a rule that Mrs. Hudson follows:

DOORBELL If the doorbell rings, believe that there is someone at the door.

It is easy to see that this rule fits the case. Mrs. Hudson's belief formation can be described as her following this conditional, whereas following the conditional means that she forms the consequent belief *because* she recognizes that the antecedent condition holds. Generalizing this, Byrne (2005, p. 94) states that

S follows the Rule R ('If conditions C obtain, believe that p') on a particular occasion iff on that occasion:

a. S believes that p because she recognizes that conditions C obtain

Which implies:

b. S recognizes (hence knows) that conditions C obtain

c. Conditions C obtain

d. S believes that p

DOORBELL happens to be a good rule, that is, it usually produces true beliefs. On the other hand, you can think of bad rules that produce false beliefs most of the time. For instance, "If you are hungry, believe that it is sunny outside" is a rule that will not generate true beliefs in general. There are simply too many instances in which you are hungry, but it is not sunny outside. In other words, the rule is unreliable.

I can now make use of epistemic rules to get a grasp of what goes on in the Otto case. The case was the following:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto,

his notebook plays the role usually played by a biological memory. I ask Otto whether he believes that the museum is on 53rd Street. Otto looks at his notebook and answers "I believe the museum is on 53rd Street."

Just as I described Mrs. Hudson's belief formation, I can now describe Otto's belief formation as a two-step process. Otto looks at his notebook, and then self-ascribes a belief because of what is written in the notebook. He recognizes that the notebook says that p, and transitions, by reasoning, to the belief that he believes that p. He thereby fits the following rule:

NOTEBOOK If your notebook says that p, believe that you believe that p.

Otto believes that he believes that p, because he recognizes that his notebook says that p. This looks quite similar to Byrne's (2005) general rule BEL: If p, believe that you believe that p (Byrne, 2005, p. 95). There is a sense in which it is just an instance of BEL, because Otto recognizes that p by looking at the notebook. However, I do not opt to use BEL here. My rationale is twofold. First, BEL shows a very strong asymmetry between a first person formulation and a third person version. We already saw that the case of extended introspection does not fit with this asymmetry to such an extent. NOTEBOOK on the other hand captures the close similarity to a third-person rule that fits our initial intuition that extended introspection is not quite as privileged and special as genuine introspection. Second, NOTEBOOK can provide additional insights to the Otto case. Both reasons will become apparent in the following section.

So far NOTEBOOK looks just like DOORBELL. However, it is more than the simple DOORBELL rule. NOTEBOOK is a very special rule, if one supposes that the conditions for extended beliefs are satisfied for Otto. Remember conditions (ii) and (iii) that external aids have to satisfy for an extended belief. The external aid has to be (ii) readily accessible, and (iii) automatically endorsed. (ii) plays an important role in making it possible to follow NOTEBOOK. Epistemic rules in general do not say whether they can actually be followed. DOORBELL, for instance, gives you a conditional that provides a transition from recognizing the doorbell ringing to a belief that will likely be true, if the antecedent holds. But you might not be able to recognize that the doorbell is ringing. You could be deaf, or simply listening to music on headphones on a volume that makes it impossible to hear the doorbell. Nothing guarantees that a good rule is one that you can follow. However, this is

different for extended believing Otto and the NOTEBOOK rule. Otto is guaranteed to be able to follow NOTEBOOK reliably with regard to recognizing the antecedent.[133] The argument for this is rather simple:

1) Otto has a belief extending to the content of his notebook. (Assumption)
2) Otto's belief can only be extended if the content of the notebook is readily accessible. (Conditions for Extended Belief)
3) Otto can readily access the content of the notebook (from 1, 2)
4) If the content of the notebook is readily accessible, then Otto can reliably recognize that the notebook says that p, if it says that p. (Spelling out Accessibility)

---

5) Otto can reliably recognize that the notebook says that p, if it says that p. (from 3, 4)

The argument shows that Otto is able to reliably recognize that the antecedent of the NOTEBOOK rule holds, if it holds. He can do so in virtue of the extended belief condition that makes the notebook readily accessible. The only step in the argument that is not an assumption or already independently argued for is (4), but I take this to be intuitively true. What else could it mean to be able to readily access a notebook, if not that I can reliably recognize that the notebook says that p, if it does say that p?[134]

That Otto can reliably recognize that the notebook says that p is not enough to guarantee that he can reliably follow NOTEBOOK completely. He further needs to be able to follow the conditional and form a belief according to the conditional. This is not worrisome at all. As long as Otto is able to reason, he is able to follow a conditional just fine.

---

[133] It is important to highlight the difference between a rule being reliable and one being able to follow the rule reliably. The rule is reliable (good) if it mostly produces true beliefs. On the other hand, one can follow an inferential rule reliably if one usually is in a position to follow it. One can reliably follow a reliable rule, but one can also reliably follow an unreliable rule. The same goes for being unable to reliably follow a rule.

[134] In earlier versions I was tempted to read the ready access condition stronger than merely reliable access. However, that would be against Clark and Chalmers (1998) intention of providing a parallel to biological memory. Clark (2010) uses reliable access instead of ready access to avoid this confusion. Thanks to Brie Gertler for pointing this out.

So far I established that Otto can reliably follow NOTEBOOK, given that he has beliefs extending to his notebook. However, I still need to provide reasons why NOTEBOOK is actually a good epistemic rule. Why should NOTEBOOK generate true rather than false beliefs? Here I take another condition of extended belief to play the pivotal role. This time it is condition (iii), the automatic endorsement condition. The idea is that whenever Otto looks into his notebook and reads that p, he automatically endorses that p and thereby is guaranteed to believe that p. This can be used in an argument as follows:

1) Otto automatically endorses that p, if he reads that p in his notebook. (Condition of Extended Belief)
2) Endorsing that p entails believing that p. (Spelling out Endorsement)
3) Otto reads that p in his notebook. (Assumption)
4) Otto endorses that p. (from 1, 3)
5) Otto believes that p. (from 2, 4)

---

6) If Otto reads that p in his notebook, he believes that p. (from 3, 4, 5)

This conclusion shows that NOTEBOOK is actually a good epistemic rule. It is good, because the consequent belief will always be true when Otto follows the rule. Whenever Otto follows NOTEBOOK, he starts by looking at the notebook which says that p. Otto recognizes that p and automatically endorses it. This endorsement guarantees that he believes that p. So when Otto follows the conditional and forms the belief that he believes that p, he will be correct. Following NOTEBOOK is infallible, because the mere act of following the rule guarantees the second-order belief to be true by securing the first-order belief.

A crucial step in the argument is (2), which depends on the notion of endorsement in play. Clark and Chalmers (1998) do not provide much information in this regard. However, Clark (2010) says that endorsement means that "It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed about as trustworthy as something retrieved clearly from biological memory" (Clark, 2010, p. 46). I take this to entail belief, insofar as it is equal to regarding p as true. If Otto recognizes that the notebook says that p, he holds p to be true without additional, critical scrutiny. He will

use p as a premise in practical and theoretical reasoning, the same as if he would hold p to be true based on any other source. This should be uncontroversial, given that I assume that the extended mind thesis is true.

Behind this argument lies a general observation of the Otto case. If Otto self-ascribes a belief by looking at the notebook two usually[135] unrelated factors coincide. On one hand there is the ground for a belief. For instance, I can form a perceptual belief based on a perceptual seeming. I form the belief that the sun is shining, because I have a visual experience of the sun shining. This is my evidence that I base my belief on. However, on the other hand there is an external fact that makes the belief true. My belief that the sun is shining is true, if the sun is in fact shining. This truthmaker is different from the basis of my belief. In the Otto case things seem different. The very same thing that Otto bases his second-order belief on also makes this second-order belief true. He looks at the notebook and believes that he believes that p because of the notebook saying that p. At the same time the notebook makes it the case that he believes that p, thereby making his second-order belief true. This makes it so difficult for Otto to be wrong about himself, if he self-ascribes by looking at the notebook. The external aid is both the basis and the truthmaker for his self-ascription. Here even the most determined sceptic cannot find a gap in which to insert his knife – as long as the sceptic is on board with extended beliefs in general.[136]

I established that NOTEBOOK is an epistemic rule that Otto can reliably follow, and moreover a good, truth-conducive rule. However, where does NOTEBOOK put extended introspection with respect to privileged and peculiar access? To answer this I want to look at third person equivalents to the NOTEBOOK rule. After all, the prior intuition was that there is nothing special about Otto looking at the notebook compared to me looking at the notebook. So perhaps there is a similar epistemic rule for me. And I believe there is, let's call it O-NOTEBOOK.

O-NOTEBOOK If Otto's notebook says that p, believe that Otto believes that p.

Under the assumption that Otto has beliefs extended to the notebook, O-NOTEBOOK also looks like a very good rule. If I look at Otto's notebook and recognize that it says that p,

[135] Even though not always. One might argue that the same thing happens if I form beliefs about my qualia.

[136] Neta (2011) discusses this feature extensively. See also Alston (1971), Chisholm (1957), and Shoemaker (1963).

then it will be true that Otto believes that p. It will be true, because by the assumption of extended beliefs whatever is written in the notebook constitutes dispositional beliefs of Otto, just the same way something stored in biological memory would. However, there are some differences to NOTEBOOK. First, O-NOTEBOOK plus the assumption that Otto has extended beliefs does not guarantee that one can reliably follow the rule. Whereas Otto can reliably follow NOTEBOOK in virtue of the extended mind conditions, there is no condition that guarantees me any access to Otto's notebook. Hence, I might not be able to recognize the antecedent of the conditional in a large number of cases.

Second, Otto can ascribe *occurrent* beliefs by using NOTEBOOK, whereas O-NOTEBOOK cannot do the same. The idea here is that whenever Otto looks at his notebook to follow his NOTEBOOK rule he thereby endorses the content right at that moment. That is, the endorsement, and thereby the belief that p, plays an active role in Otto's cognitive machinery at the moment of him following NOTEBOOK. Moreover, it is plausible that Otto will be consciously aware of his belief that p, when he follows NOTEBOOK. On the other hand, if I look at Otto's notebook, there is no way for me to tell whether Otto believes that p occurrently or dispositionally. It is possible that I ascribe to Otto the belief that p by looking at his notebook, while at that moment Otto himself does not look at the notebook at all and is thinking about something completely unrelated to p. In this case I can still correctly ascribe a belief that p to Otto, but only a dispositional belief. The difference can be spelled out by expanding both epistemic rules:

NOTEBOOK* If your notebook says that p, believe that you occurrently believe that p.

O-NOTEBOOK* If Otto's notebook says that p, believe that Otto occurrently believes that p.

NOTEBOOK* is a good rule, whereas O-NOTEBOOK* is not. The former will produce mostly (always) true beliefs, but the latter generates a ton of false beliefs in cases where I follow O-NOTEBOOK* when Otto does not look at his notebook. Hence there is a difference between NOTEBOOK and O-NOTEBOOK insofar as they both use a general notion of 'belief', but have different types of beliefs as truthmakers in general.

With these differences in mind I can confidently say that there is something peculiar and special about extended introspection. But it is only a minor difference. Nothing guarantees that I can reliably follow the third person equivalent to Otto's NOTEBOOK rule. Moreover, even when I can follow O-NOTEBOOK, I cannot employ quite the same method as Otto. However, I can do something in the vicinity, closely resembling Otto's belief formation. And I can be reliable as well; I am just limited in the range of reliable extended mental state ascriptions. I cannot reliably ascribe occurrent states based on Otto's external aids, but I can reliably attribute his extended beliefs in general. The result is somewhere in between the features that ordinary self-knowledge and mind-reading are said to possess. It is not quite as peculiar as self-knowledge based on usual introspection, but there is still some difference between the first-personal access and the third-personal one.

Using epistemic rules to model extended introspection might prompt the question whether we would not be better off with epistemic rules for self-knowledge in general. Should we just take on board Byrne's (2018) overall story instead of the single process model and add rules for extended beliefs as a special case of BEL? One reason to resist this concession is to emphasize the goal of solving the standing state problem discussed in chapter 3. The single process model is in a better position to explain how one can know the standing states one has right now without worrying about changing them whenever one attempts to know what mental state one is in. A second reason to hold on to the single process model is that it seems to fit into a cognitive story that has independent support. And third, one might argue that the epistemic rules for extended states are too different to Byrne's other rules to make them a unified account. Byrne's rules are meant to explain privileged and peculiar access, but the NOTEBOOK rule does not provide the same extent of privilege and peculiarity. Therefore, it should not be grouped together with rules like BEL. However, these reasons are not decisive. It is unclear to what extent different epistemic rules should and should not be grouped together. I do not have a convincing argument against grouping NOTEBOOK together with BEL in a single account. Moreover, whether the single process model ultimately proposes the right cognitive story is up for debate and hopefully will be answered by scientists in the future. It might turn out that it is completely off-track then we should not hold on to it at all. Finally, one might propose that the standing state problem is not all that important anyway. In chapter 3 I hinted at responses to the problem following this line of reasoning. For instance, Gertler (2011a) proposed that Moran (2001) should explicitly endorse that the only relevant self-knowledge is knowledge of one's mental states

after introspecting. Given this route one does not take the standing state problem as something to be solved. In this case Byrne's proposal does look very attractive and more unified than the single process model. However, I explicitly aimed to provide a transparency account that does avoid the standing state problem and hence ought to hold on to the single process model. This comes with the cost of treating extended mental states differently than non-extended mental states.

7.6 Conclusion

I showed that a straightforward interpretation of extended introspection as genuine introspection is not appropriate. Hence, the single process model cannot account knowledge of all one's mental states. Nevertheless, this cannot be used as an objection against the single process model. Knowledge of extended states is not in the same way privileged or special that knowledge of non-extended mental states is. Otto's extended belief is accessible by me in a similar way that it is accessible to Otto himself. On the other hand, there is an asymmetry between Otto's access to his non-extended belief and my access to Otto's non-extended belief.

Furthermore, a self-directed mind-reading story does not fit either. This left us with four distinct options. We can change our accounts of introspection and mind-reading, deny the extended-mind thesis, or propose a sui generis form of extended introspection. I chose the latter. Therefore, I proposed an original account of extended introspection based on epistemic rules. I provided a rule for the Otto example, and argued that this happens to be a rule that Otto can follow and that generates true beliefs. Both are secured by the requirements of extended beliefs. If Otto has extended beliefs, then he can reliably follow the NOTEBOOK rule which leads him to true beliefs about his mental states. Finally, I showed how this picture fits the ideas of privileged and peculiar access. Otto is privileged, because having beliefs extended to the notebook (plus reasoning) guarantees that he can reliably follow the NOTEBOOK rule. Furthermore, he is in a special position because he can attribute that a belief is occurrent based on NOTEBOOK, whereas a third-personal variation of the rule, O-NOTEBOOK, cannot do the same. It is still an open question to what extent this result can generalize to other cognitively integrated artifacts. This is mainly a question of whether (i)-(iii) hold for all extended mental states, or whether some artifact can be integrated enough to count as part of the mind without satisfying (i)-(iii). Plausibly, widespread technology such as smartphones and smartwatches fit these conditions.

8 Conclusion

Throughout the thesis I developed a proposal for a unified transparency account of self-knowledge. The single process model provides an explanation for the asymmetry between beliefs about my own mental states and beliefs about other people's mental states. Usually, when we form a mental state we also generate a corresponding, dispositional second-order belief about that state. And when the situation allows – or requires – this dispositional belief can become occurrent. My aim was to see how much explanatory power this idea has. Can we explain the peculiar nature of *all* self-knowledge with the single process model? Almost. The model can be applied to all types of mental states, except extended mental states. However, knowledge of one's extended mental states is different from knowledge of one's non-extended mental states. Extended mental states do not show the strong asymmetry in my access to other people's access to them. Hence, I argued that knowledge of extended mental states is not genuine self-knowledge, and excluding those from the single process model is a palatable result.

Furthermore, the single process model accounts for the privileged access one has to one's own mental states with leaving enough room for fallibility. We can know our own mental states, but we are not perfect in finding out what state we are in at a given moment. Sometimes we just get things wrong about ourselves. We can be biased, lack concepts, or simply be the victim to the malfunction of a neural process

The path towards the single process model followed three steps. In chapters 1 to 3 I laid the groundwork discussing the preferred definition of the explanandum self-knowledge. I argued that the peculiar nature of self-knowledge has to be described on the level of belief and not on the level of language. Moreover, I explained what exactly I aim for when I propose a *transparency* account of self-knowledge, and why it is not obvious how such an account can be *unified*, that is, how it can work for all types of mental states. In chapters 4 and 5 I developed the single process model as an explanation of the features of self-knowledge laid out earlier. I showed how my proposal can achieve unification and transparency. Chapter 6 provided a cognitive story that fits my proposal. I showed how the predictive processing framework can tell a story about the mechanisms that are responsible for the principles proposed in the single process model. This story is still in its first steps, but shows that the assumptions made in the single process model can fit into a bigger picture of human cognition. In chapter 7 I conceded that extended mental states are also

part of this picture, but knowledge of these states cannot be explained with the single process model. I argued that this is no reason to worry, because we should rather think of these states as incompatible with any form of classical introspection. Extended introspection does not show the same features that self-knowledge of merely internal states has. Hence, knowledge of extended mental states is not the target phenomenon I want to explain with the single process model. Instead, we should accept a distinct way of getting knowledge about these extended states.

What I provided in this thesis is the basis of an original account of self-knowledge. However, there is still work to be done. On more than one occasion I referred to various scientific fields, such as psychology, neuroscience, or biology. The single process model is an empiricist proposal that at some points relies on non-philosophical explanations. I take it that it is ultimately not the job of an armchair philosopher to explain the concrete neurological processes for self-knowledge. Hence, when I referred to mental state forming processes the content of these terms has to be filled in by research from other fields. Similarly, the predictive processing story I proposed will most likely only be a first starting point and change in accordance with empirical research in the near future. The intended upshot of my proposal is not that self-knowledge works exactly like I describe here – though it would be great if it turns out to be fully correct. Rather, I want to bring a new kind of explanation into the discussion. An explanation based on the idea that first-order mental state formation and second-order belief formations usually occur together. The whole thesis is an argument that this option should be taken seriously. The idea can explain the peculiar nature of self-knowledge and is compatible with our understanding of our cognition. I am sure there are other – perhaps better – ways of spelling out the idea, but the most important step is to offer the idea In the first place.

Bibliography

Adams, F. R., & Aizawa, K. (2010). Defending the bounds of cognition. In R. Menary (Ed.), *The Extended Mind* (pp. 67-80). Cambridge, MA: MIT Press.

Alston, W. (1971). Varieties of Privileged Access. *American Philosophical Quarterly, 8*, pp. 223–241.

Anscombe, G. E. M. (1957). *Intention.* Oxford: Basil Blackwell.

Armstrong, D. M. (1968). *A Materialist Theory of Mind.* London: Routledge & Kegan Paul.

Ashwell, L. (2013a). Deep, dark...or transparent? Knowing our desires. *Philosophical Studies, 165*(1), pp. 245-256.

Ashwell, L. (2013b). *Review of Transparent Minds: A Study of Self-Knowledge*. Retrieved February 11, 2017, from Notre Dame Philosophical Reviews: http://ndpr.nd.edu/news/42227-transparent-minds-a-study-of-self-knowledge/

Audi, R. (1994). Dispositional Beliefs and Dispositions to Believe. *Noûs, 28*(4), pp. 419-434.

Baillargeon, R. (1994a). How Do Infants Learn About the Physical World? *Current Directions in Psychological Science, 3*(5), pp. 133-140.

Baillargeon, R. (1994b). Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology, 12*(1), pp. 9-33.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*(91), pp. 612-625.

Bar-On, D. (2004). *Speaking My Mind.* Oxford: Oxford University Press.

Bar-On, D. (2010). Avowals: Expression, Security, and Knowledge:Reply to Matthew Boyle, David Rosenthal, and Maura Tumulty. *Acta Analytica, 25*, pp. 47-63.

Bar-On, D. (2011). Neo-Expressivism: Avowals' Security and Privileged Self-Knowledge. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 189-201). Oxford: Oxford University Press.

Bar-On, D., & Sias, J. (2013). Varieties of Expressivism. *Philosophy Compass, 8*(8), pp. 699-713.

Bartlett, G. (2017). Occurrent States. *Canadian Journal of Philosophy, 48*(1), pp. 1-17.

Bayne, T., & Montague, M. (2011). Cognitive Phenomenology: An Introduction. In T. Bayne, & M. Montague (Eds.), *Cognitive Phenomenology* (pp. 1-34). New York: Oxford University Press.

Ben-Yami, H. (1997). Against Characterizing Mental States As Propositional Attitudes. *Philosophical Quarterly, 47*(186), pp. 84-89.

Berker, S. (2008). Luminosity Regained. *Philosophers' Imprint, 8*(2), pp. 1-22.

Bernecker, S. (2010). *Memory: A Philosophical Study.* Oxford: Oxford University Press.

Bernecker, S. (2014). How to Understand the Extended Mind. *Philosophical Issues*(24), pp. 1-23.

Bilgrami, A. (2006). *Self-Knowledge and Resentment.* Cambridge, MA: Harvard University Press.

Blanchard, C., Roll, R., Roll, J.-P., & Kavounoudias, A. (2011). Combined contribution of tactile and propriocetive feedback to hand movement perception. *Brain Research*(1382), pp. 219-229.

Block, N. (1990). Inverted Earth. *Philosophical Perspectives, 4*, pp. 53-79.

Boring, E. G. (1953). A History of Introspection. *Psychological Bulletin, 50*(3), pp. 169-189.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eye see. *Nature*(391), p. 756.

Boyle, M. (2009). Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research, LXXVIII*(1), pp. 133-164.

Boyle, M. (2011). Transparent Self-Knowledge. *Proceedings of the Aristotelian Society Supplementary, LXXXV*, pp. 223-241.

Brentano, F. (1995 (1874)). *Psychology from an empirical standpoint* (2. ed.). (A. C. Rancurello, D. B. Terrell, & L. L. McAlister, Trans.) New York: Routlage.

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention and motor preparation. *Frontiers in Psychology, 2*(218).

Brown, J. (2008). Subject-Sensitive Invariantism and the Knowledge Norm for Practical Reasoning. *Noûs, 42*(2), pp. 167-189.

Bruce, V., Green, P., & Georgeson, M. (2003). *Visual Perception: Physiology, Psychology, & Ecology.* Hove: Psychology Press.

Brueckner, A. (2011). Neo-Expressivism. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 170-188). Oxford: Oxford University Press.

Bruner, J. S. (1990). *Acts of Meaning.* Cambridge, MA: Harvard University Press.

Burge, T. (1988). Individualism and Self-Knowledge. *The Journal of Philosophy, 85*(1), pp. 649-663.

Burge, T. (1996). Our Entitlement to Self-Knowledge I. *Proceedings of the Aristotelian Society, New Series, 96*, pp. 91-116.

Byrne, A. (2005). Introspection. *Philosophical Topics, 33*(1), pp. 79-104.

Byrne, A. (2011). Transparency, Belief, Intention. *Proceedings of the Aristotelian Society Supplementary Volume, LXXXV*, pp. 201-220.

Byrne, A. (2012a). Knowing What I See. In D. Smithies, & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 183-209). Oxford: Oxford University Press.

Byrne, A. (2012b). Knowing What I Want. In J. Liu, & J. Perry (Eds.), *Consciousness and the Self: New Essays* (pp. 165-183). Cambridge: Cambridge University Press.

Byrne, A. (2018). *Transparency and Self-Knowledge.* Oxford: Oxford University Press.

Carruthers, P. (2010). Introspection: Divided and Partly Eliminated. *Philosophy and Phenomenological Research, LXXX*(1), pp. 76-111.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press.

Carter, A. J., & Pritchard, D. (Forthcoming). Extended Self-Knowledge. In J. Kirsch, & P. Pedrini (Eds.), *Third-Person Self-Knowledge, Self-Interpretation, and Narrative.* Berlin: Springer.

Cassam, Q. (2011). Knowing What You Believe. *Proceedings of the Aristotelian Society, CXI*, pp. 1-23.

Cassam, Q. (2014). *Self-Knowledge for Humans.* Oxford: Oxford University Press.

Cassam, Q. (2017). What asymmetry? Knowledge of self, knowledge of others, and the inferentialist challenge. *Synthese*(194), pp. 723-741.

Chisholm, R. (1957). *Perceiving: A Philosophical Study.* Ithaca, NY: Cornell University Press.

Churchland, P. M. (1988). Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor. *Philosophy of Science*(55), pp. 167-187.

Churchland, P. M. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.* Cambridge: Cambridge University Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* Oxford: Oxford University Press.

Clark, A. (2010). Memento's Revenge: The Extended Mind, Extended. In R. Menary (Ed.), *The Extended Mind* (pp. 43-66). Cambridge, MA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), pp. 1-24.

Clark, A. (2016). *Surfing Uncertainty.* Oxford: Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), pp. 7-19.

Conee, E., & Feldman, R. (1998). The Generality Problem for Reliabilism. *Philosophical Studies*(89), pp. 1–29.

Das, N. (2018). *Śrīharṣa*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy (Spring 2018 Edition): https://plato.stanford.edu/archives/spr2018/entries/sriharsa/

Davies, M., & Gardner, D. (2010). *Frequency Dictionary of American English.* New York: Routledge.

Dennett, D. C. (2013). Expecting ourselves to expect: The Bayesian brain as a prejector. *Behavioral and Brain Sciences, 36*(3), pp. 29-30.

DeRose, K. (2002). Review of Knowledge and Its Limits by Timothy Williamson. *The British Society for the Philosophy of Science, 53*(4), pp. 573-577.

Descartes, R. (2008 (1641)). *Meditations on First Philosophy.* (M. Moriarty, Ed.) Oxford: Oxford University Press.

Dewhurst, J. (2017). Folk Psychology and the Bayesian Brian. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group.

Douven, I. (2006). Assertion, Knowledge, and Rational Credibility. *The Philosophical Review, 115*(4), pp. 449-485.

Downey, A. (2017). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*.

Drummer, T., Picot-Annand, A., Neal, T., & Moore, C. (2009). Perception. *Movement and the rubber hand illusion*(38), pp. 271-280.

Dudai, Y. (2004). The Neurobiology of Consolidations, or How Stabke is the Engram? *Annual Review of Psychology*(55), pp. 51-86.

Edgley, R. (1969). *Reason in Theory and Practice.* London: Hutchinson.

Eliasmith, C. (2005). A New Perspective on Representational Problems. *Journal of Cognitive Science*(6), pp. 97-123.

English, H. B. (1921). In Aid of Introspection. *American Journal of Psychology*(32), pp. 404-417.

Evans, G. (1982). *The varieties of reference.* (J. McDowell, Ed.) Oxford: Oxford University Press.

Fernández, J. (2013). *Transparent Minds: A Study of Self-Knowledge.* Oxford: Oxford University Press.

Finkelstein, D. (2003). *Expression and the Inner.* Cambridge, MA: Harvard University Press.

Finkelstein, D. (2012). From Transparency to Expressivism. In G. Abel, & J. Conant (Eds.), *Rethinking Epistemology* (Vol. 2, pp. 101-118). Berlin: De Gruyter.

Frankfurt, H. (1988). Identification and Wholeheartedness. In H. Frankfurt, *The Importance of What We Care About* (pp. 159-176). Cambridge: Cambridge University Press.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences*, pp. 181-197.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Coputational Biology, 4*(11).

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), pp. 127-138.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*(151).

Gallois, A. (1996). *The World Without, the Mind Within: An Essay on First-Person Authority.* Cambridge: Cambridge University Press.

Gazzaniga, M. S. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga, *The Cognitive Neurosciences.* Cambridge, MA: MIT Press.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric. *Brain*(123), pp. 1293-1326.

Gerken, M. (2013). The Roles of Knowledge Ascriptions in Epistemic Assessment. *European Journal of Philosophy, 23*(1), pp. 141-161.

Gerken, M. (2017). *On Folk Epistemology: How we Think and Talk about Knowledge.* Oxford: Oxford University Press.

Gertler, B. (2007). Overextending the mind? In B. Gertler, & L. Shapiro (Eds.), *Arguing About the Mind* (pp. 192-206). New York: Routledge.

Gertler, B. (2011a). *Self-Knowledge.* New York: Routledge.

Gertler, B. (2011b). Self-Knowledge and the Transparency of Belief. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 125-145). Oxford: Oxford University Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), pp. 559–582.

Goldman, A. (1970). *A theory of human action.* Princeton: Princeton University Press.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press.

Goldman, A. (2008 (1979)). What Is Justified Belief? In E. Sosa, J. Kim, J. Fantl, & M. McGrath, *Epistemology: An Anthology* (2. ed., pp. 333-347). Blackwell.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts and Theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1994). The Theory Theory. In L. Hirschfield, & S. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257-93). New York: Cambridge University Press.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanism and the Theory Theory. *Psychological Bulletin, 138*(6), pp. 1085-1108.

Grzankowski, A. (2012). Not All Attitudes are Propositional. *European Journal of Philosophy, 23*(3), pp. 374–391 .

Guerraz, M., Provost, S., Narison, R., Brugnon, A., Virolle, S., & Bresciani, J.-P. (2012). Integration of Visual and Proprioceptive Afferents in Kinesthesia. *Neuroscience*(223), pp. 258-268.

Harman, G. (1990). The Intrinsic Quality of Experience. *Philosophical Perspectives, 4*, pp. 31-52.

Hawthorne, J., & Stanley, J. (2008). Knowledge and Action. *The Journal of Philosophy, 105*(10), pp. 571-590.

Heersmink, R. (2015). Dimensions of Integration in Embedded and Extended Cognitive Systems. *Phenomenology and the Cognitive Sciences, 13*(3), pp. 577-598.

Held, R. (1965). Plasticity in sensory-motor systems. *Scientific American*, pp. 84-94.

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems 6.* San Mateo, CA: Morgan Kaufmann.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*(268), pp. 1158-1160.

Hohwy, J. (2013). *The Predictive Mind.* Oxford: Oxford University Press.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), pp. 687-701.

Horgan, T., & Kriegel, U. (2007). Phenomenal Epistemology: What is consciousness that we may kniw it so well? *Philosophical Issues*(17), pp. 123-144.

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*(436), pp. 71-77.

Husserl, E. (1991 (1907-09)). *On the phenomenology of the consciousness of internal time (1893-1917). Husserliana 10.* (J. Brough , Trans.) Dordrecht: Kluwer.

Hutto, D. D. (2008). *Folk Psychological Narratives: The Socioculteral Basis of Understanding Reasons.* Cambridge, MA: MIT Press.

Jackson, F. (1982, April). Epiphenomenal Qualia. *The Philosophical Quarterly, 32*(127), pp. 127-136.

James, W. (1890). *Principles of Psychology.* New York: Holt.

Jensen, M. S., Yao, R., Street, W. N., & Simons, D. J. (2011). Change Blindness and Inattentional Blindness. *WIREs Cognitive Science, 2*, pp. 529-546.

Jongepier, F., & Strijbos, D. (2015). Introduction: self-knowledge in perspective. *Philosophical Explorations, 18*(2), pp. 123-133.

Kawaro, M., Hayakama, H., & Inui, T. (1993). A Forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*(4), pp. 415-422.

Kloosterboer, N. (2015). Transparent Emotions? *Philosophical Explorations, 18*(2), pp. 246-258.

Krueger, J. (2012). Seeing Mind in Action. *Phenomenology and the Cognitive Sciences, 11*(2), pp. 149-173.

Kvanvig, J. (2009). Assertion, Knowledge, and Lotteries. In P. Greenough, & D. Pritchard (Eds.), *Williamson on Knowledge* (pp. 140-160). Oxford: Oxford University Press.

Lackey, J. (2008). What Should We Do When We Disagree? In T. S. Gendler, & J. Hawthorne (Eds.), *Oxford Studies in Epistemology* (pp. 274-293). Oxford: Oxford University Press.

Lackey, J. (2010). A justificationist view of disagreement's epistemic significance. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology* (pp. 298-325). Oxford: Oxford University Press.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America*.

Lishman, J. R., & Lee, D. N. (1973). The autonomy of visual kinaesthesis. *Perception, 2*, pp. 287-294.

Locke, J. (1975 (1689)). *An Essay Concerning Human Understanding.* (P. H. Nidditch, Ed.) Oxford: Oxford University Press.

Lycan, W. (1987). *Consciousness.* Cambridge, MA: MIT Press/A Bradford Book.

Lycan, W. (1996). *Consciousness and Experience.* Cambridge, MA: MIT Press.

Macdonald, C. (2014). In my 'Mind's Eye': introspectionism, detectivism, and the basis of authoritative self-knowledge. *Synthese, 191*(15), pp. 3685-3710.

Mack, A., & Rock, I. (1998). *Inattentional Blindness.* Cambridge, MA: MIT Press.

Maravita, A., Spence, C., & Driver, J. (2003). Multisensory Integration of the Body Schema: Close to Hand and Within Reach. *Current Biology, 13*, pp. R531-R539.

McGlynn, A. (2014). *Knowledge First?* Basingstoke: Palgrave Macmillan.

Menary, R. (2007). *Cognitive Integration: Mind and Cogntition Unbounded.* Basingstoke: Palgrave Macmillan.

Menary, R. (2010). Dimensions of Mind. *Phenomenology and the Cognitive Sciences, 9*(4), pp. 561-578.

Moore, G. E. (1903). The Refutation of Idealism. *Mind, 12*(48), pp. 433-453.

Moore, G. E. (1951). Russell's Theory of Descriptions. In P. A. Schilpp (Ed.), *The Philosophy of Bertrand Russell: Library of Living Philosophers* (pp. 175-225). New York: Tudor.

Moran, R. (2001). *Authority and estrangement.* Princeton: Princeton University Press.

Moran, R. (2003). Responses to O'Brien and Shoemaker. *European Journal of Philosophy, 11*(3), pp. 402-419.

Neta, R. (2011). The Nature and Reach of Privileged Access. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 9-32). New York: Oxford University Press.

Nichols, S., & Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford: Oxford University Press.

Nisbett, R., & Wilson, T. (1977). Telling More Than We Can Know: Verbal Reports On Mental Processes. *Psychological Review, 84*(3), pp. 231-259.

O'Brien, L. (2003). Moran on Agency and Self-knowledge. *European Journal of Philosophy*(11), pp. 375-390.

O'Shaughnessy, B. (2000). *Consciousness and the World.* Oxford: Oxford University Press.

Parent, T. (2007). Infallibilism about self-knowledge. *Philosophical Studies, 133*(3), pp. 411-424.

Parent, T. (2016). The Empirical Case against Infallibilism. *Review of Philosophy and Psychology, 7*(1), pp. 223-242.

Peacocke, A. (2017). Embedded mental action in self-attribution of belief. *Philosophical Studies, 174*(2), pp. 353-377.

Peacocke, C. (1998). Conscious Attitudes, Attention, and Self-Knowledge. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing Our Own Minds.* Oxford: Oxford University Press.

Pettigrew, R. (2015). Pluralism About Belief States. *89*(1), pp. 187-204.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*(134), pp. 17-35.

Pitt, D. (2004). The Phenomenology of Cognition or "What Is It like to Think That P?". *Philosophy and Phenomenological Research, 69*(1), pp. 1-36.

Pritchard, D. (2005). *Epistemic Luck.* Oxford: Clarendon Press.

Pritchard, D. (2007). Epistemic Luck. *Synthese*(158), pp. 277-298.

Pryor, J. (2005). There is immediate justification. In M. Steup, & E. Sosa (Eds.), *Contemporary debates in epistemology* (pp. 181-202). Oxford: Blackwell.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), pp. 79-87.

Reed, B. (2002). How to Think about Fallibilism. *Philosophical Studies, 107*(2), pp. 143-157.

Reynolds, R. F., & Bronstein, A. M. (2003). The broken escalator phenomenon. *Experimental Brain Research*(151), pp. 301-308.

Rodriguez-Ortiz, C. J., & Bermúdez-Rattoni, F. (2007). Memory Reconsolidation or Updating Consolidation? In F. Bermúdez-Rattoni (Ed.), *Neural Plasticity and Memory: From Genes to Brain Imaging.* Boca Raton, FL: CRC Press.

Roessler, J. (2013). The silence of self-knowledge. *Philosophical Explorations, 16*(1), pp. 1-17.

Rosenthal, D. (1986). Two Concepts of Consciousness. *Philosophical Studies, 94*(3), pp. 329-359.

Rupert, R. (2004). Challenges to the Hypothesis of Extended Cognition. *Journal of Philosophy*(101), pp. 343-356.

Ryle, G. (1984 (1949)). *The Concept of Mind.* Chicago: University of Chicago Press.

Samoilova, K. (2016). Transparency and introspective unification. *Synthese, 193*(10), pp. 3363–3381.

Sara, S. J. (2000). Retrieval and Reconsolidation: Toward a Neurobiology of Remembering. *Learning & Memory*(7), pp. 73-84.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review, 117*(2), pp. 245-273.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience, 5*(2), pp. 97-118.

Setiya, K. (2011). Knowledge of intention. In A. Ford, J. Hornsby, & F. Stoutland (Eds.), *Essays on Anscombe's intention* (pp. 170-197). Cambridge, MA: Harvard University Press.

Shah, N., & Velleman, D. J. (2005). Doxastic Deliberation. *The Philosophical Review, 114*(4), pp. 497-534.

Shoemaker, S. (1963). *Self-Knowledge and Self-Identity.* Ithaca, NY: Cornell University Press.

Shoemaker, S. (1982). The Inverted Spectrum. *Journal of Philosophy, 79*(7), pp. 357-381.

Shoemaker, S. (1994). Self-Knowledge and "Inner Sense": Lecture II: The Broad Perceptual Model. *Philosophy and Phenomenological Research, 54*(2), pp. 271-290.

Silins, N. (2013). Introspection and Inference. *Philosophical Studies*(163), pp. 291-315.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception, 28*, pp. 1059-1074.

Smithies, D. (2012a). A Simple Theory of Introspection. In D. Smithies, & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 259-293). Oxford: Oxford University Press.

Smithies, D. (2012b). Mentalism and Epistemic Transparency. *Australasian Journal of Philosophy, 90*(4), pp. 723-741.

Smithies, D. (2016). Belief and Self-Knowledge: Lessons From Moore's Paradox. *Philosophical Issues, 26*(1), pp. 393-421.

Snowdon, P. (2012). How to Think about Phenomenal Self-Knowledge. In A. Coliva, *The Self and Self-Knowledge* (pp. 243-262). Oxford: Oxford University Press.

Spaulding, S. (2015). On Direct Social Perception. *Consciousness and Cognition*(36), pp. 472-482.

Srinivasan, A. (2015). Are We Luminous? *Philosophy and Phenomenological Research, XC*(2), pp. 294-319.

Stefanucci, J. K., & Proffitt, D. R. (2008). Skating down a steeper slope: Fear influences the perception of geographical slant. *Perception*(37), pp. 321-323.

Stefanucci, J. K., & Proffitt, D. R. (2009). The Roles of Altitude and Fear in the Perception of Height. *Journal of Experimental Psychology*(35), pp. 424-438.

Sterelny, K. (2004). Externalism, Epistemic Artefacts and the Extended Mind. In R. Schantz (Ed.), *The Externalist Challenge* (pp. 239-254). Berlin: De Gruyter.

Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), pp. 465-481.

Stokes, D. (2013). Cognitive Penetrability of Perception. *Philosophy Compass, 8*(7), pp. 646-663.

Strawson, G. (2015). Self-Intimation. *Phenomenology and the Cognitive Sciences, 14*(1), pp. 1-31.

Sutton, J. (2006). Distributed cognitions:dimains and dimensions. *Pragmatics and Cognition, 14*(2), pp. 235-247.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The Psychology of Memory, Extended Cognition, and Socially Distributed Remembering. *Phenomenology and the Cognitive Sciences, 9*(4), pp. 521-560.

Thagard, P. (2006). Desires are not propositional attitudes. *Dialogue, 45*(1), pp. 151-156.

Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering application.* New York: D. Van Nostrand.

Tye, M. (2000). *Consciousness, Color and Content.* Cambridge, Mass: MIT Press.

van Schalkwyk, G. I., Volkmar, F. R., & Corlett, P. R. (2017). A predictive Coding Account of Psychotic Symptoms in Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*.

van Ulzen, N., Semin, G. R., Oudejans, R. R., & Beek, P. J. (2008). Affective stimulus properties influence size perception and the Ebbinghaus illusion. *Psychological Research*(72), pp. 304-310.

Vega, J. A. (2007). Thinking About Me, You, and Them: Understanding Higher-Order Propositional Attitudes. *The New School Psychology Bulletin, 5*(1), pp. 75-105.

Velleman, J. D. (1992). The Guise of the Good. *Noûs, 26*(1), pp. 3-26.

Vierkant, T. (2015). How do you know that you settled a question? *Philosophical Explorations, 18*(2), pp. 199-211.

Vogel, J. (2010). Luminosity and Indiscriminability. *Philosophical Perspectives, 24*(1), pp. 547-572.

von Helmholtz, H. (1860 (1962)). *Handbuch der physiologischen Optik* (transl. & ed., Vol. 3). Dover: J. P. C. Southall.

Weatherson, B. (2004). Luminous Margins. *Australasian Journal of Philosophy, 82*(3), pp. 373-383.

Wegner, D., & Wheatley, T. (1999). Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist*(54), pp. 480-492.

Weilnhammer, V., Stuke, H., Hesselmann, G., Sterzer, P., & Schmack, K. (2017). A Predictive Coding Account of Bistable Perception - a Model-Based fMRI Study. *PLoS Computational Biology, 13*(5), p. e1005536.

Wiese, W. (2017). What are the contents of representations. *Phenomenology and the Cognitive Sciences, 16*(4), pp. 715–736.

Williamson, T. (2000). *Knowledge and its Limits.* Oxford: Oxford University Press.

Wilson, T. (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, MA: Harvard University Press.

Wittgenstein, L. (2009). *Philosophical Investigations (Translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte)* (4th ed.). Chichester: Wiley-Blackwell.

Wright, C. (1989). Wittgenstein's Later Philosophy Of Mind: Sensation, Privacy, and Intention. *Journal of Philosophy*(86), pp. 622-643.

Wright, C. (1998). Self-Knowledge: The Wittgensteinian Legacy. In C. Wright, B. Smith, & C. Macdonald, *Knowing our Minds* (pp. 13-45). Oxford: Oxford University Press.

Wright, C. (2001). *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations.* Cambridge, MA: Harvard University Press.

Wright, C. (2015). Self-knowledge: the reality of privileged access. In S. C. Goldberg (Ed.), *Externalism, Self-Knowledge, and Skepticism* (pp. 49-74). Cambridge: Cambridge University Press.

Wundt, W. (1888). Selbstbeobachtung und innere Wahrnehmung. *Philosophische Studien*(4), pp. 292-309.